AL/OE-TR-1995-0190

THE COORDINATION OF DYNAMIC VISUAL AND AUDITORY SPATIAL PERCEPTS AND RESPONSIVE MOTOR ACTIONS: A REVIEW AND INTEGRATION OF CURRENT THEORY AND RESEARCH

Bartholomew Elias

OCCUPATIONAL AND ENVIRONMENTAL HEALTH DIRECTORATE
Bioenvironmental Engineering Division
Noise Effects Branch
2610 7th Street
Wright-Patterson AFB OH 45433-7901

ARMSTRONG LABORATORY

December 1995

Interim Technical Report for the Period October 1994-September 1995

Approved for public release; distribution is unlimited.

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Armstrong Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Rd., STE 0944
Ft. Belvoir VA 22060-6218

TECHNICAL REVIEW AND APPROVAL

AL/OE-TR-1994

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

Robert A. Lee
ROBERT A. LEE
Chief, Noise Effects Branch

FOR THE DIRECTOR

James D Montgomery
JAMES D. MONTGOMERY, LT COL, USAF, BSC
Chief
Bioenvironmental Engineering Division
Armstrong Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
December 1995

3. REPORT TYPE AND DATES COVERED Interim Report for the Period October 1994 - September 1995

4. TITLE AND SUBTITLE
The Coordination of Dynamic Visual and Auditory Spatial Percepts and Responsive Motor Actions: A Review and Integration of Current Theory and Research

5. FUNDING NUMBERS

PE:	62202F
PR:	ILIR
TA:	23000B
WU:	23000B51

6. AUTHOR(S)

Bartholomew Elias

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Armstrong Laboratory, Occupational & Environmental Health Directorate
Bioenvironmental Engineering Division
Human Systems Center
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7901

8. PERFORMING ORGANIZATION REPORT NUMBER

AL/OE-TR-1995-0190

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This review is intended to serve as an integration of current research and thought regarding visual and auditory spatial perception in an effort to consolidate diverse theoretical viewpoints and empirical findings. The aim of this technical report is to derive a unified account of bimodal spatial perception and responsive motor actions. The model of bimodal information processing developed in this report will serve as a basis for defining issues of concern regarding the implementation of bimodal spatial displays that require the individual to integrate dynamic spatial information acquired both aurally and visually. This model draws upon theory and research from ecological and information processing perspectives of auditory and visual perception. The model posits that information regarding the spatial layout and spatial dynamics of the environment is incrementally accrued through multiple modalities including vision and audition to form a common functional representation of spatio-temporal parameters. The key parameters of dynamic spatial position and motion conveyed in this functional representation are subsequently utilized to formulate plans and programs for responsive actions. Currently, a research program examining the process of cross-modal integration of dynamic auditory and visual information is underway. The findings of this research will have important implications for the implementation of 3-D auditory display technology for conveying dynamic spatial information in high visual workload environments.

14. SUBJECT TERMS

Visual Perception
Motor Control
Auditory Perception

15. NUMBER OF PAGES
150

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

TABLE OF CONTENTS iii

LIST OF FIGURES v

LIST OF TABLES vii

PREFACE viii

INTRODUCTION 1

DYNAMIC SPATIAL COORDINATION 12

The Coordination of Perception and Action 12

The Environment, Objects, and Events 13

The Perceiver 16

VISUAL SPATIALIZATION 20

Parallel Processing of Form and Motion 20

Spatial Filtering, Edge Detection, and Form Segregation 23

Motion Detection 27

Dynamic Cues for Depth Perception 43

Static Cues for Depth Perception 56

Stereoscopic Vision 61

AUDITORY SPATIALIZATION 67

The Perception of Acoustical Structure 67

Duplex Theory of Auditory Localization 69

The Role of Anatomical Transfer Functions 75

Static Auditory Acuity 77

Auditory Motion Detection 79

Auditory Depth Perception 82

SPATIAL CORRESPONDENCE BETWEEN VISION AND AUDITION 89

A Common Spatial Metric 89

Visual Cues for Auditory Spatialization 91

Auditory Cues for Visual Spatialization 93

Bimodal Facilitation 97

CONTROL OF MOTOR ACTIONS 100

Rapid Aimed Movements 100

Continuous Manual Control 108

Locomotion 116

AN INTEGRATED THEORY OF SPATIO-TEMPORAL DYNAMICS 119

REFERENCES 124

LIST OF FIGURES

Figure		Page
1.	(a) Size constancy from texture occlusion (b) The invariant horizon ratio	15
2.	(a) Outflow of textural features (b) Inflow of textural features (c) Lateral flow of textural features	18
3.	Spatial frequency and spatial phase properties of X and Y cells	21
4.	Spatially tuned filters of the magnocellular and parvocellular pathways	23
5.	A lattice of directionally specific phase modulation detectors	32
6.	The Ternus display	35
7.	Examples of the adjacency principle	37
8.	Petersik's (1989) framework for motion analysis	41
9.	A schematic of the rate of optical expansion created by convergence of objects with the viewpoint	46
10.	(a) In-phase oscillation of stimulus edges (b) Anti-phase oscillation of stimulus edges	48
11.	Locally tuned changing size channels	49
12.	A model of motion in depth processing	51
13.	(a) Differential motion of random 2-D patterns to illicit dynamic 3-D percepts (b) Orthographic projections of counter-rotating cylinders	54
14.	The correspondence problem in stereoscopic vision	62
15.	Activity of neurons sensitive to crossed, uncrossed, and near-zero disparities	64
16.	The interaural path difference	72

Figure		Page
17.	The auditory cone of confusion	74
18.	Stochastic estimation of sound source position and velocity	81
19.	Stimuli used in O'Leary and Rhodes (1984)	98
20.	(a) A Test-Operate-Test-Exit (TOTE) unit (b) A series of TOTE units comprising a motor action	109
21.	A modified version of the optimal control model (OCM)	112
22.	A model of multimodal sensory integration	120

LIST OF TABLES

Table Page

1. Principal characteristics of the "short-range" and "long-range" motion processes 40

THIS PAGE INTENTIONALLY LEFT BLANK

PREFACE

This report contains a literature review of research on visual and auditory spatial integration completed under the USAF Armstrong Laboratory In-house Laboratory Independent Research program (WU 23000B51).

This review is based in part on a review of the literature that the author completed in partial fulfillment for the requirements for his doctoral degree in Psychology from the Georgia Institute of Technology. The author is grateful for the critical review of that material by his preliminary examination and dissertation committee members and their encouragement to pursue this topic in greater detail. The author acknowledges these members of his committee, namely Dr. Elizabeth Davis, Dr. Dennis Folds, Dr. M. Jackson Marr, Dr. Neff Walker, and his advisor, Dr. Gregory M. Corso. The author also acknowledges the resources provided by the Noise Effects Branch of Armstrong Laboratory (AL/OEBN) in helping to make this effort possible.

THIS PAGE INTENTIONALLY LEFT BLANK

INTRODUCTION

The ability to maneuver and emit motor acts directed at environmental stimuli is fundamentally dependent on a coordination between multimodal sensory inputs and between these integrated inputs and subsequent motor responses. Moreover, this coordinative process must be relatively continuous to deal with the complex spatial dynamics of the environment. James Gibson (see, e.g., Gibson, 1979) posited that examination of these spatial dynamics must focus upon the continuous flow of information available through the spatio-temporal patterning of dynamic physical events. (see, e.g., Gibson, 1979) focused upon the dynamics of the visual scene, however many of his followers have further extended these views to the domain of auditory perception (see, e.g., Schiff & Oldak, 1990; Shaw, McGowan & Turvey, 1991). The impact of this ecological approach to perception lies in its provision for a concise description of environmental objects and events of importance to the perceiver through spatio-temporally defined invariants or consistencies. According to this approach, these invariant properties in the environment provide important cues or, in Gibson's terms, *affordances*, regarding objects and events. In this sense, Gibson's notion of *invariants* and *affordances* provides a direct, coordinated linkage between sensation and action. Furthermore, this approach stresses the important role of a dynamic perceiver that acts upon its surrounding environment in order to observe the emergent invariant relationships the exist between objects and their surrounds. While Gibson's theory introduces the notion of the perceiver as a dynamic element within the environment, his theory fails to be an adequate account of the complex biological processes involved in perception. It is well established that perceivers transduce, analyze, and synthesize perceptual information and subsequently emit responsive motor acts within a intricate coordination of physiological systems. Therefore, it is evident that a consideration of multimodal perceptual integration and motor coordination is incomplete without the inclusion of an information-processing description of the perceiver.

In viewing the perceiver as an information-processing organism, one must examine the various modalities, pathways, and codes that are implemented in the processing of information that impinges upon the sensory organs. In analyzing information about spatial arrangements and spatial dynamics of the external environment, two modalities of particular interest are vision and audition. The specific encoding, analysis, and synthesis of spatial information in these two modalities is explored in detail. Moreover, the integration of information analyzed by the visual and auditory modalities to produce a unified spatial percept is explored. Finally, the role of these integrated dynamic percepts in guiding complex motor actions such as rapid aimed movements, continuous tracking, and locomotion is considered.

In visual perception, light reflected from distal environmental objects is projected onto the human perceiver's two retinas creating two slightly eccentric two-dimensional proximal images. The light cast upon these two retinal mosaics of photoreceptors is thereupon transduced into patterns of neural activity. These patterns are partially determined by the nature of the photoreceptors' sensitivity to the intensity and wavelength of light, but are also reflective of the subsequent structure of neural connections in the visual system. The most significant of these neural connections is the early segmentation of information into two separate parallel pathways, the magnocellular and parvocellular pathways (Livingstone & Hubel, 1988; Lennie, et al., 1990). These distinct pathways are presumed to carry separate visual codes, the magnocellular being responsible for conveying low-resolution dynamic position and motion information, whereas the parvocellular is predominantly responsible for conveying high-resolution form information. Moreover, luminance contrast information is presumed to be segmented early in the visual system through patterns of neural activity across a wide range of spatial frequency selective channels. Only low spatial frequency information is presumed to be carried through the magnocellular pathway, whereas, detailed high spatial-frequency information is coded in the parvocellular pathway.

The transient, low spatial frequency tuning of the magnocellular system renders it important in conveying information about rapid, temporally dynamic patterns such as motion and flicker. Thus, information analysis conducted along this pathway has been described as giving rise to a "short-range" percept of motion (Braddick, 1974; Braddick, 1980). However, this "short-range" percept of motion can be ambiguous due to the *correspondence problem* that arises because of inherent uncertainty regarding which points in two time-sampled proximal images refer to the same physical features in a dynamic environment. Resolution of this ambiguity relies upon top-down constraints regarding rules of object dynamics as well as integration of form information from parvocellular processing in order to converge upon a common "long-range" motion solution. Based on psychophysical observations, it is posited that the dynamic spatial layout is incrementally built up from interactions of the magnocellular dominated "short-range" motion process and the subsequent analysis of form mediated predominantly by the parvocellular pathway and the imposition of featural constraints comprising the "long-range" motion detection process. While the theoretical distinction between “short-range” and “long- range” motion solutions sufficiently accounts for motion along the fronto-parallel plane, it fails to provide a complete account of three-dimensional visual spatial perception since it ignores the processing of motion-in-depth information and the perception of changing optical flow patterns associated with locomotion of the perceiver.

When objects in the environment move in depth or when the perceiver locomotes through the environment, local and global changes in the flow patterns of the optic array occur. These flow patterns induce expansion and compression of feature boundaries projected in the proximal images. Local expansion or compression suggests motion of an environmental object in depth, whereas global expansion or compression is indicative of locomotion by the perceiver. David Lee (1980) suggested that the relative rate of local or global optical expansion uniquely defines motion-in-depth. Moreover, the variable specifying this relative rate of expansion, signified by the Greek letter tau (τ), further specifies the time-to-contact of environmental objects with the point of observation and is therefore critical for collision avoidance during locomotion. Psychophysical and physiological research conducted by David Regan and his colleagues (see Regan, Beverley & Cynader, 1979 for an overview) has demonstrated a biological basis for the computational processing of this motion-in-depth information within the visual system. In Regan, et al.'s (1979) model of processing motion-in-depth information, a solution for the recovery of dynamic three-dimensional object motion is developed through several stages. The optical flow pattern is sufficiently specified monocularly, but in the case of local changes, it is ambiguous with respect to changing size or changing position in depth. Optic flow information can be disambiguated through the accumulation of other monocular cues such as texture gradients and occlusion when such information is sufficiently defined in the environment. Moreover, as with fronto-parallel motion, a complete motion-in-depth solution must incorporate top down constraints and three-dimensional stereoscopic form information. In this sense, optic flow may be regarded as giving rise to a short-range or low level solution for motion-in-depth. However, a long-range or high level solution is requisite of time-sampled monocular and binocular depth cues.

The specification of monocular and binocular cues for depth and their interactions has been a topic of empirical inquiry for some time. Of particular interest is the manner in which these depth cues such as occlusion, relative size, foreshortening, linear perspective, shading, and binocular disparity are combined to produce a singular percept of object depth. The combined influence of monocular pictorial cues and binocular disparity information for disambiguating the percept of depth is viewed as a cooperative additive factor process (Landy, et al., 1991). However, utilization of these cues appears to be dependent upon prior form segregation and therefore suggests that such a solution is derived relatively late in the course of visual processing . Moreover, computation of binocular disparity requires comparisons of visual codes from the two eyes and therefore, it too occurs relatively late in the course of visual processing. Not surprisingly, both monocular pictorial cues and binocular disparity are presumed to be mediated predominantly through parvocellular pathways and are assumed to be constituents of a high level or long-range computation of depth.

From these theoretical accounts of visual processing based upon voluminous research literature recounting psychophysical and physiological findings, it can be posited that the visual system reconstructs a representation of the three-dimensional dynamic spatial layout through an incremental accrual of motion and form information. The manner in which this representation is incrementally accrued can be likened to a series of cascading filters converging upon a single cohesive array of spatio-temporal parameters. In this manner, rapid motion information can be accrued and augmented by more detailed form and position information obtained through monocular and binocular image filtering and synthesis in which three-dimensional object structure and position is recovered.

Spatial information can also be accrued through the auditory modality. However, due to the inherent non-spatial character of acoustic signals, spatial information must be recovered through a concomitant analysis of acoustical form. Consequently, unlike vision, in audition the analysis of location and motion is interdependent with an analysis of acoustical structure. In auditory perception, perturbations of air caused by environmental sound sources impinge upon receptive cells on the basilar membrane in the inner ears where transduction into neural codes signaling pitch and loudness occurs. Three-dimensional localization information is reflected in the characteristics of these neural codes and can be understood through a consideration of the perturbations that the intensity and frequency spectrum characteristics of the sound signal undergo prior to reaching the two inner ears (see Middlebrooks & Green, 1991 for a recent review). These perturbations are created by features of the environment as well as features of the listener's anatomy. Absorption and diffraction of the sound wave by the listener's head leads to interaural intensity, time and phase differences that can be analyzed through time-sampled neural analysis of the sound profile. While these interaural differences can provide information regarding azimuthal sound source location, adequate determination of sound source location can only be understood by consideration of the distortion cues provided by the structure of the pinna or outer ear (see Batteau, 1967). This filtering by the pinna significantly alters the sound spectra of high frequency sources and consequently can act as a monaural cue to sound source localization particularly important for detecting spatial location in elevation. While the pinna only provides localization cues at relatively high frequencies, at low frequencies, diffraction of sound waves by the listener's upper torso can potentially provide cues to location. The role of these various anatomical distortions can be described through head-related transfer functions (HRTFs) that describe the perturbation of sound waves by the listener's anatomical features.

Prior the contact with the listener, sound emanating from an object can be altered as it travels through the air and is reflected and absorbed by materials in the environment. The alterations of the frequency spectrum produced by these perturbations of the sound wave as it travels over distance to reach the listener provide auditory cues to depth (cf., Coleman, 1963; Mershon & King, 1975). These cues are provided by the decrement in sound pressure level or intensity over distance, the differential attenuation of frequency components as they travel through air, and the differential absorption and reflectance of frequency components by materials in the acoustical environment. These cues give rise to percepts reflecting the egocentric distance of sound sources. Similarly, acoustical properties of dynamic sound sources such as the *Doppler shift* and the relative rate of change in sound level can provide information regarding their relative location and motion in depth. Thus, distortions in the spectral profiles of environmental sound sources by structures in the environment and structures of the listener's anatomy provide cues from which a representation of the dynamic spatial layout can be acoustically derived.

However, auditory and visual spatial percepts rarely are derived independently. Rather, these spatial representations appear to be conveyed through a common spatial metric or yardstick. Consequently, comparisons across modalities regarding spatial layout can be made without reliance on translational processes (cf., Auerbach & Sperling, 1974). Nonetheless, the predominance of vision in conveying information about spatial layout biases the weighting of spatial cues in favor of those obtained visually. This visual dominance can result in the aiding of auditory spatial judgments, but can also bias spatial percepts to favor visual information over auditory information in the presence of cross-modal discrepancy (see Welsh & Warren, 1986 for a review). However, the visual modality suffers from a restriction in the field-of-view and consequently is incapable of providing continuous spatial information across the full 360^{o} in azimuth and elevation. Since the auditory system is not spatially restricted in the same manner as the visual system, it can provide a means for conveying dynamic spatial information regarding objects and features that lie outside the immediate field of view. For example, Perrott and his colleagues (see Perrott, et al., 1990, Perrott, et al., 1991) have demonstrated that peripheral auditory spatial cues can aid visual search performance. Moreover, the use of localized auditory display information has been shown to aid in the task of air traffic avoidance under visual flight conditions (see Begault, 1993). Thus, preliminary research indicates that auditory spatial cues can serve to aid in the detection of visual objects under conditions imposed by visually demanding task scenarios.

However, the interaction of multimodal representations is dependent on temporal as well as spatial contiguity. In fact, the spatio-temporal characteristics of information presented in one modality has a significant impact on the perception of concomitant signals presented through a second modality (O'Leary & Rhodes, 1984). In general, temporal covariance of multimodal signals has been repeatedly shown to provide redundancy gain in signal detection tasks (see, e.g., Loveless, Brebner, & Hamilton, 1970). Redundancy gain from bimodal spatial and temporal covariants in signal presentation remains to be demonstrated, however evidence strongly suggests that vision and audition are commonly conveyed through a unified spatio-temporal representation. The additive and interactive contributions of multimodal cues to this spatio-temporal representation provide a means through which redundancy gain can be observed as a function of bimodal redundancy. Furthermore, this common spatio-temporal representation is presumed to be the vehicle through which plans for motor action are formulated in reference to the dynamic spatial layout of the environment.

Complex motor activities consist of sequences of rapid, ballistic movements arranged in hierarchical temporal sequences (cf., Miller, Galanter, & Pribram, 1960; Pew & Rosenbaum, 1988). Thus, higher level motor action plans serve to specify the spatial and temporal parameters for specific motor programs that signal the execution of discrete motor actions. These spatial and temporal parameters are perceptually specified by the incremental accrual of information regarding the three-dimensional dynamic spatial layout of the performance environment and provide the requisite information for motor guidance. Mass properties of the organism and objects in the environment are parameterized through learning and the prior history of the organism and are specified in motor programs along with spatio-temporal information regarding the dynamic layout of the environment. This learning occurs through the provision of knowledge of results regarding the outcomes of guided motor actions thereby allowing for precise specification of appropriate mass, spatial and temporal parameterization of future actions (cf., Salmoni, Schmidt, & Walter, 1984). These space, time, and mass parameters serve to computationally specify the force required for ballistic motor responses as well as the synchronization of action sequences (cf., Schmidt, 1975).

Thus, rapid ballistic aimed movements of particular extremities serve as the functional units of more complex motor activities. Lawful function relationships between the spatio-temporal demands of these motions and the speed and accuracy of execution have been demonstrated in tasks that are spatially and/or temporally constrained. Furthermore, research has suggested that the nature of these relationships between spatio-temporal demands and movement speed and accuracy are mediated by the microstructure of submovements as well as by task demands. The microstructure of submovements was originally posited to involve iterative corrections based upon sensory feedback regarding the discrepancy between actual and desired positions of the extremity in relation to the target of the motor action. However, more recent theories have suggested that these submovements are the result of an optimizing strategy that attempts to dynamically re-specify force parameters in the presence of inherent neuromotor noise (cf., Meyer, et al., 1988; Meyer, et al., 1990). Thus, the speed and accuracy characteristics of a single motor action unit is functionally dependent on the quality of initial spatio-temporal response parameters and the degree of central and peripheral variability in the transmittal of these parameters through the psychomotor pathway. In spatially constrained tasks, the size and distance of the target at which aimed movements are made adequately reflect task difficulty and consequently serve as sufficient parameters for describing motor performance. In temporally constrained tasks, on the other hand, movement velocity directly influences endpoint variability in aimed movements and consequently it serves as a sufficient parameter for describing task performance in such situations. In the case of indirect manipulation tasks, the control gain and control order of the device used to mediate and transmit motoric responses to the action environment interact to determine overall task difficulty. Thus, it is evident that the programming of motor responses involves the parameterization of spatio-temporal characteristics of the environment and mass and force properties of the organism's effectors and manipulation devices associated with intermediating control systems. Furthermore, the specification of these parameters is stochastically optimized to counteract the effects of inherent noise in the central and peripheral psychomotor pathways.

While most research examining discrete motor actions limit the degrees of freedom of motion to one axis, it has been demonstrated that these general relationships between the speed and accuracy of movement and spatio-temporal characteristics of the task environment are extensible to motor actions made along two and three-dimensions of control movement (see, e.g., Jagacinski & Monk, 1985). In these multi-axis movement tasks, it has been demonstrated that the coordination of effectors in relationship to these axes of motion is not completely independent, nor is it completely integrated. That is, movement trajectories along multiple axes exhibit characteristics that indicate neither a pure Euclidean metric nor a pure city-block metric for parameterizing multi-axis motor coordination. These characteristics of multi-axis control performance as a function of the distance and size of the movement target have been demonstrated to hold for motions involving the head and eyes as well as the extremities indicating the ubiquity of such controlling relationships.

When these simple ballistic units of motor actions are temporally and hierarchically sequenced they functionally operate as coordinated components of a motor planning sequence. In dynamic task environments, the execution of such plans require continuous monitoring of outcomes in relation to desired results and this feedback serves to update force parameters for subsequent motor responding. In continuous manual control of dynamic systems, the continuous updating of spatio-temporal force parameters can be described through a feedback loop such as that detailed in the optimal control model (OCM) of iterative corrections regarding misalignments between actual and desired system states (cf., Baron & Levinson, 1980). These iterative feedback loops serve to update state space parameters in the internal spatio-temporal representation of the task environment and dynamically correct for sensori-motor delays and transmittal noise during the course of information processing. Consequently, these parameters optimally specify the force parameters required for corrective motor responses. In this manner, perceptual-motor coordination can be achieved in highly dynamic environments.

The ability to specify spatial and temporal parameters for continual motor responding in these dynamic task environments is dependent upon the quantity and quality of state information that can be conveyed through sensory channels. The quantity and quality of this information can be increased through augmentation of the display that conveys system state space information to the operator. Such augmentation can be achieved by providing increased preview of upcoming system states and spatio-temporal representations of predicted future conditions (cf., Poulton, 1974). Such augmentation can be conveyed through pursuit and compensatory displays in either the visual or auditory modality. System states are typically conveyed through visual displays, however auditory frequency, intensity, and spatial cues have also been used for providing information regarding the relative discrepancy between actual and desired system states. The quality of control responding to information conveyed through these alternative modalities is functionally dependent on the perceptual abilities of the operator to detect changes in the properties of the tracking display whether they be spatial or featural in nature.

In tasks involving more direct manipulation of the environment, such as locomotion, spatio-temporal characteristics of the terrain and potential obstacles are similarly represented though a build-up of multimodal sensory information. These spatio-temporal parameters serve to define the requirements for action plans regarding responsive movements. Moreover, the continual updating of relative changes in these spatio-temporal parameters conveyed through acoustical and optical flow serve to iteratively update the relationships between the dynamic organism and features of the surrounding environment. In this manner, spatio-temporal characteristics of the environment can be continually updated through acoustical and optical flow parameters such as the point of expansion (POE) which defines the direction of travel and tau (τ) which specifies the relative rate of travel and the instantaneous time-to-contact with objects in the environment. Through the incorporation of these parameters into motor response plans, the organism can successfully navigate through a cluttered terrain and emit responsive actions in reference to environmental features. In this manner, a complex coordination between multimodal sensory inputs and motor responses can be achieved.

The continual coordination of multimodal sensory percepts and motor actions is fundamentally dependent on a process in which spatio-temporal characteristics of the environment can be parameterized through a successive build-up of information conveyed by multiple modalities and pathways. A theoretical model is developed in which spatio-temporal information comes to be conveyed through a common representation of the dynamic spatial layout that can receive cascading information from various auditory and visual pathways. The cascading nature of information flow in this model allows for the incremental accrual of increasingly detailed motion, form, and location parameters from successively higher levels of processing within a respective modality. In this manner, spatio-temporal parameters for motor plans can be estimated expeditiously thereby allowing for rapid responding to potentially threatening conditions in the environment and can be incrementally updated and optimized through the accrual of fine detail spatial information thereby allowing for precision in motor responding. Such a scheme in which auditory and visual spatio-temporal information is conveyed through a combined, weighted representation of the dynamic spatial layout makes several predictions regarding the utility of bimodal cueing in complex perceptual-motor domains. In particular, it is expected that the degree of spatio-temporal congruity between auditory and visual inputs should directly influence task performance. Moreover, the incremental build-up of information from multiple modalities into this common representation of the dynamic spatial layout suggests that the auditory modality, which is not as spatially restricted as the visual modality, can provide for early estimation of spatio-temporal parameters specified in plans for motor action. It is posited that, to the extent auditory and visual spatio-temporal information can be equated and commonly represented, auditory position and motion information at peripheral locations can be employed as preview for visually guided motor tasks. This auditory preview information can consequently aid performance in motor tasks directed at visual objects to the extent that it is compatible or veridical with respect to the position and motion of its corresponding visual object (see Elias, 1994).

DYNAMIC SPATIAL COORDINATION

The Coordination of Perception and Action

The nature of an organism's ability to integrate multimodal sensory inputs and emit appropriate motor actions within the complex dynamics of the physical environment is a phenomenon of particular interest in experimental psychology. The scientific inquiry into such processes becomes quickly engrossed in the complex details of specific sensory modalities on the one hand and precise descriptions of motoric actions on the other. Far too infrequently are there any systematic attempts to integrate the functional processes of multimodal sensation and complex motor responses into unified theoretical accounts. Therefore, in attempting to describe the complex coordination of multimodal sensation and motor performance, the thesis presented by John Dewey in his classic essay, *The Reflex Arc Concept in Psychology*, (Dewey, 1896, reprinted in Watson, 1979) is particularly apropos. In criticizing the disjoint analysis of sensory processing and motor responding prevalent in the experimental methodologies and theoretical accounts of perception and action during his time, Dewey asserted that:

> ...the reflex arc idea, as commonly employed, is defective in that it assumes sensory stimulus and motor response as distinct psychical existences, while in reality they are always inside a coordination and have their significance purely from the part played in maintaining or reconstituting the coordination... It is the coordination which unifies that which the reflex arc concept gives us only in disjointed fragments. (Dewey in Watson, p. 233 and 240).

This notion of a coordination between perception and action has been recapitulated in Pew and Rosenbaum's (1988) recent review of research on human motor control. Pew and Rosenbaum (1988) refer to the difficulty in empirically and theoretically addressing the nature of this coordination as the *perceptual-motor integration problem.* In their review, the nature of this problem is discussed in reference to motor performance research on continuous manual control and rapid discrete movements. Clearly, Dewey's comments are relevant to understanding many of these current research trends in the study of human performance. For example, in manual tracking tasks, the coordination between human motor responses to system disturbances and outcomes is typically represented in terms of functional models that describe this coordination between the human and the machine (see e.g., Poulton, 1974; Sheridan & Ferrell, 1974; Wickens, 1984). Similarly, Fitts' Law and similar accounts of rapid aimed motor movements focus on the functional relationships between visual objects and motor acts directed toward them (e.g., Schmidt, et. al, 1979; Meyer, et al., 1990). These functional descriptions emphasize the coordination between qualities of the stimulus and aspects of the motor response. However, these research paradigms can become prone to criticisms like those levied by Dewey (1896) when they fail to address the simple fact that motor acts directed at stimuli necessarily change the nature of those stimuli. As Dewey (1896, in Watson, 1979, p. 239) emphasizes, "the so-called response is not merely *to* the stimulus; it is *into* it". Clearly, the tracking and rapid motor response paradigms attempt to account for the inherent change in the stimulus introduced by the response, however, more satisfactory theoretical accounts of response consistency in the face of this ever present change can be obtained from studies in ecological psychology that assume an active perceiving organism. Thus, an exploration of the nature of interactions between objects, events, and motor responses made by an active perceiver from the standpoint of ecological psychology provides a common description of the spatial layout and spatial dynamics of the environment that can serve as a solid foundation for theoretical accounts of sensori-motor coordination.

The Environment, Objects, and Events

This notion of consistency of action in the presence of a seemingly ever-changing world or environment was a major impetus for the ecological approach to perception and action often referred to as the Gibsonian view after its principle advocate, James Gibson (see Gibson, 1979). Gibson (1979) introduced the concept of stimulus invariants, or functional relationships of stimulus properties that remain consistent over time and space. Consequently, invariants can only be defined spatio-temporally, and are revealed to the observer or organism through active exploration of the environment which includes locomotion and the emission of motor acts directed toward environmental stimuli. Gibson therefore maintains that the important features for perception are revealed in the dynamic flow and disturbances in the ambient optic array defining the totality of light reaching the observer (cf., Bruce & Green, 1985) . While Gibson stressed the importance of the visual features of the environment, his principles can be generally extended to the realm of auditory perception. Indeed, as Gibson suggests that the ambient optic array serves to define the information utilized by the visual modality, the ambient acoustic properties of an environment serve to define the information utilized by the auditory modality (cf., Handel, 1989; Folds, 1990). Presently, detailed consideration will be given to visual objects and events whose perception arises from spatio-temporal sampling of the optic array while auditory ambience will be considered later within the context of auditory spatial perception.

Gibson (1979) considers three major focal areas in visual spatial perception: (1) the perception of surface layout, (2) the perception of changes in that layout, and (3) the perception of movement on behalf of the perceiver. The first of these areas revolves around the perception of objects in the environment, whereas the latter two areas focus on the perception of events (cf., Johansson, von Hofsten, & Jansson, 1980). In the visual pickup of object information, Gibson (1979) stresses the importance of structure. However, Gibson does not consider object structure in the classical psychophysical analysis of form which ignores the importance of figure-ground relationships to focus instead on object form devoid from the role of the stimulus surrounds. On the contrary, Gibson contends that the perception of objects must be considered through the invariant structural relationships that exist between objects in the environment and the background texture elements from which they are segregated. Consequently, Gibson termed his theory a ground theory of perception and distinguishes it from the more prevalent air theories that analyze stimuli in isolation (cf., Bruce & Green, 1985).

Invariant structural relationships can be represented in static scenes through higher-order relationships or ratios between objects and their surrounds. For example, in the perception of three dimensional layout, depth is conveyed in the environment through the changing texture density gradient. That is, as depth increases, the size and spacing between texture elements decreases. This texture density information provides the background upon which object size constancy emerges. In particular, an object's size is conveyed by the number of texture elements it occludes (see Figure 1a) . Therefore, size is not an inherent visual property of the object but rather is an emergent invariant that describes the object in relation to its environment. The structure of the visual environment can also provide the perceiver with knowledge of his own position and orientation. For example, the horizon provides important invariant information conveyed in the horizon ratio relation which is consistent across environmental objects identical in height. The ratio of the proportion of upright objects above and below the horizon line is constant regardless of distance (see Figure 1b). Hence, sampling of these horizon ratio relations can provide sufficient information for determining eye-height above the terrain (Sedgewick, 1986). Through perception of such invariant properties relationships between the perceiver and environmental features can be directly ascertained.

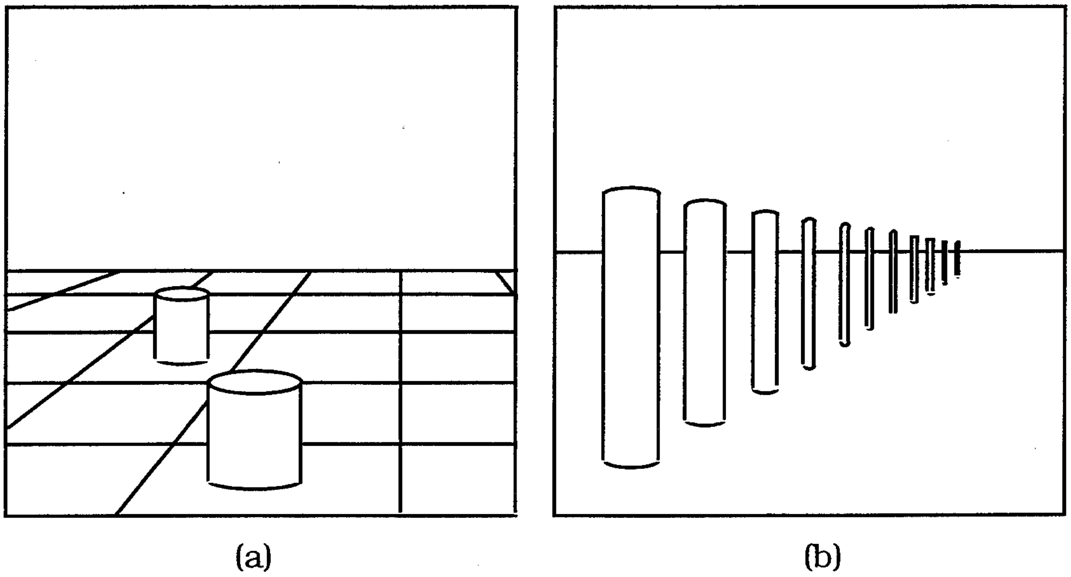

Figure 1. Examples of (a) size constancy from texture occlusion and (b) the invariant horizon ratio of environmental objects providing information regarding eye height (after Gibson, 1979).

However, Gibson asserted that cases involving static observation of an unchanging environment by the perceiver such as those discussed above are quite uncommon. Rather, objects in the environment including the perceiver exhibit motion and change. Therefore, while Gibson recognizes the role of static pictorial cues, he focuses primarily on emergent higher-order invariants in the dynamic optical array. These changes in the optic array over time serve to identify events. In particular, local disturbances in the optic array specify movement of objects in the environment whereas global disturbances in the optic array specify movement of the perceiver.

Thus, motion of objects in three-dimensional space parallel to the line of sight can be conveyed by the progressive covering and uncovering of textural elements as an object translates across the terrain. Motion in depth is similarly conveyed by this reversible occlusion of textural features, but is further specified by the rate of magnification or minification of the object (Gibson, 1979; see also Lee, 1980; Bruce & Green, 1985; Warren, 1988). This relative rate of magnification or minification can be uniquely specified by a single spatio-temporal invariant specifying the rate of motion in depth or conversely specifying the time before the object in motion will collide with the viewpoint (see Lee, 1980). This specification of motion in depth from invariant characteristics of the pattern of magnification and minification will be considered in greater detail in Section 2.4 on dynamic cues for depth perception. In the current discussion, the relevance of such an invariant is readily apparent for collision avoidance with dynamic objects in the environment and therefore is of significant survival value. The notion of magnification and minification also points to the role of invariants in distinguishing motion from changing size. Ambiguities regarding whether minification or magnification signals changes in object size or changes in object depth are particularly problematic to "air" theories of perception. However, Gibson's "ground" theory provides a direct means for disambiguating motion from changing size by specifying changes in the object relative to its surround. That is, while changing size is specified solely by magnification or minification, motion in depth is specified by magnification or minification coupled with reversible occlusion of textural features. Hence, assumptions of object rigidity are unnecessary for specifying object motion since object form is uniquely specified through segregation of the object from the surrounding textural elements. Consequently, elastic transformations of the object coupled with motion in three-dimensions is directly perceived through the spatial continuity and reversible nature of textural occlusion. Thus, three dimensional motion of fluid objects can be directly perceived by their local magnification, minification and the systematic covering and uncovering of textural features upon which these objects move.

The Perceiver

While local changes and textural occlusion specify motion of external objects, global motion of the optic array uniquely specifies the parameters of observer motion in three dimensions. As the perceiver locomotes about the environment, the global changes in the optic array form a motion perspective that uniquely specify the direction and speed of travel (Gibson, 1979; Warren, 1988; Larish & Flach, 1990; Mestre, 1992). Locomotion in depth, perpendicular to the direction of gaze, is specified by inflow and outflow of the optic array (see Figure 2). Outflow indicates forward motion or approach. The center of the outflow pattern, termed the point of expansion, specifies the destination of travel on the current trajectory. Conversely, inflow indicates backward motion or recession from the central point of convergence. The point of convergence serves to specify the perceiver's origin or the point at which the direction of travel last changed. Furthermore, shifts in the point of expansion or point of convergence signal changes in direction, whereas preservation of the same point of expansion or point of convergence over time specifies motion along a single, constant trajectory. A constantly changing point of expansion or point of convergence corresponds to a continuous turn, and consequently in such cases these parameters only specify the instantaneous direction of motion. Nonetheless, the rate of turn can be sufficiently conveyed through the invariant rate of optical flow with respect to fronto-parallel and depth axes (see Lee, 1980; Mestre, 1992). Finally, motion parallel to the direction of gaze produces global optical flow in the direction opposite the perceiver's motion (see Figure 2c). In all three of these cases, the velocity or speed of travel is directly conveyed in the rate of change in the optical flow field. The rate of motion produced by optical flow is functionally specified by the velocity of the perceiver and the perceiver's height above the terrain (Lee, 1980; Larish & Flach, 1990) . The manner in which eye-height above the terrain can be determined through the invariant horizon ratio has be previously discussed. Thus, given that both eye-height above the terrain and the rate of global optical flow can be optically specified, absolute velocity of the perceiver can be derived from optical cues and can provide information for the guidance of egomotion. Similarly, motion parallax produced by lateral egomotion, parallel to the line of sight, produces optical flow that is specified by the velocity of motion as well as the height above the terrain and the distance of environmental objects. Since egocentric distance, eye-height, and the rate of lateral optical flow can be specified by invariant relationships, absolute velocity parallel to the line of sight can similarly be optically determined. More detailed consideration of how such rate of change variables signal the speed of motion will be provided in later discussions of the computational issues surrounding motion perception along the fronto-parallel plane and motion in depth.

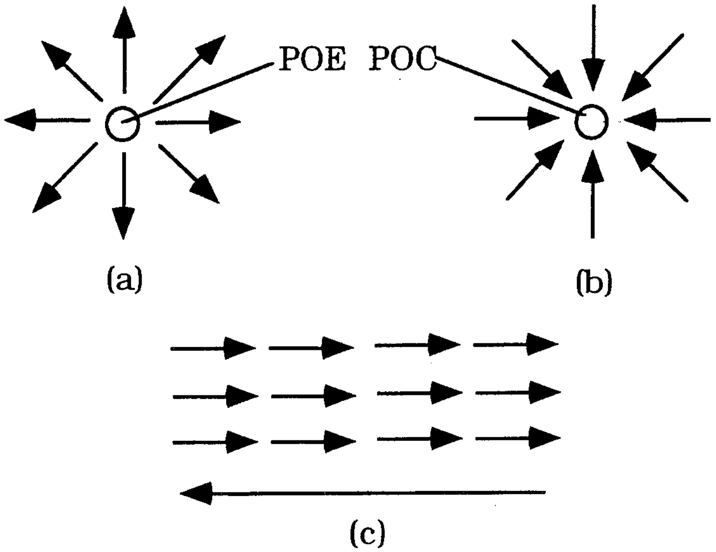

Figure 2. Locomotion of the perceiver specified by (a) *outflow* indicated forward motion or approach, (b) *inflow* specifying backward motion or recession, and (c) lateral flow of the optic array opposite the direction of travel. In locomotion in depth (a and b) the point of expansion (POE) and point of convergence (POC) uniquely define the destination and origin of travel on the current trajectory.

Thus, according to Gibson, invariant properties of objects and the environment provide information about the spatial layout of the environment and the occurrence of events within that environment such as object motion and locomotion on behalf of the perceiver. Gibson also describes how this specification can provide information regarding action. This specification for action is conveyed through affordances that specify what a particular object or event offers, provides, or furnishes for the perceiving organism (Gibson, 1979). Based upon the utility of a given affordance given the organism's particular status at the given moment in time, appropriate actions can be made. Consequently, "...the concept of affordances provides a powerful way to bridge the gap that exists in more cognitive theories between 'perception' and 'action' " (Bruce & Green, 1985, p. 199). Indeed, Gibson's theory in totality serves to describe the coordination of perception and action within the complex dynamics of the environment. However, while Gibson's theory focuses on specifying the relationship between the perceiver and its environment, it fails to provide an adequate description of the computational processing that takes place within the perceiving organism. By suggesting that the linkage between perception and action is direct, Gibson fails to address the intricate physiology and psychology of the organism that mediates the coordination between perception and action. Nonetheless, his theory is important for several reasons. First, it stresses the role of higher-order variables or invariants critical for perception and action. Second, it focuses upon the perception of dynamics and the specification of these dynamic properties through higher-order invariants. Finally, it stresses the importance of texture and ground in conveying information about objects and events in the environment. Consequently, Gibson's theory provides an important framework that specifies the variables critical to the perception of three dimensional spatial layout and spatial dynamics. As Gibson stresses, the emergence of such percepts is not the result of the analysis of static, isolated object features, but rather is the result of higher-order structural and dynamic relationships between objects and their surrounds. These specifications provide the framework upon which an integrated computational account of perception and action can be based. That is, while Gibson ends his inquiry with consideration of perceptually important environmental relationships, a complete account of the organism must go beyond this environmental description and focus upon the processing of this information through neural channels from initial transduction of the optic array to the emission of a responsive motor action. In the following chapters, theoretical accounts of these computational processes that mediate the relationship between multimodal sensory inputs regarding three-dimensional spatial dynamics on the one hand and guided motor responses on the other will be considered in detail. Physiological and psychophysical evidence for these computational descriptions will be discussed and these theoretical and empirical accounts of multimodal sensation and motor actions will be integrated into a cohesive account of dynamic spatial perception and action guidance. First, building upon Gibson's framework for visual perception, detailed consideration will be given to computational models of visual spatialization.

VISUAL SPATIALIZATION

The perception of spatial location in the visual modality is the result of an analysis of the proximal images cast upon the retinal mosaics of the two eyes. Early segmentation of the neural codes representing these proximal images carries separable information through two parallel pathways, the magnocellular and parvocellular pathways. These two pathways can be clearly distinguished by the spatio-temporal tuning and receptive field sizes of component neurons. These distinguishing features suggest that the magnocellular system is particularly adapted to detecting motion whereas the parvocellular system is responsible for, among other things, form or pattern perception. Theoretical accounts suggest that the magnocellular system may act independently to produce rapid "short-range" solution to changing position, whereas more complete "long-range" motion solutions arise from cooperation between the magnocellular and parvocellular systems. The contribution of the parvocellular system lies in its identification of elemental and textural features that provide the basis upon which invariant structural relationships can be computed. Moreover, the parvocellular system is posited to mediate the computation of binocular disparity. Consequently, the parvocellular system can be seen as contributing a large number of static viewer-centered depth cues which can be time sampled and combined with dynamic motion cues computed through the magnocellular system to produce a complete neural representation of the spatial layout and spatio-temporal dynamics of the environment.

Parallel Processing of Form and Location

Upon transduction of the optic array of light impinging upon the retinal mosaics of the two eyes, a representation of this optical structure comes to be conveyed through neural codes. As the photoreceptive rod and cone cells transmit information to the bipolar cells and subsequently to the retinal ganglion cells forming the optic nerve, these signals are functionally segregated into two distinct parallel pathways: the magnocellular and parvocellular pathways (cf., Livingstone & Hubel, 1988; Lennie, et. al., 1990). These pathways are comprised of cells that can be distinguished by their receptive field sizes and spatio-temporal response profiles. That is, while cells of the magnocellular system typically exhibit large receptive fields and demonstrate low spatial frequency tuning and high temporal frequency tuning, cells of the parvocellular system have smaller receptive fields and demonstrate tuning to higher spatial frequencies and little or no responding to temporal frequency. Consequently, cells of the magnocellular system have been described as being transient or phasic, whereas cells of the parvocellular system have been described as being sustained or tonic (cf., Lennie, et al., 1990). These pathways have further been described as giving rise to information regarding the "where" and "what" of the dynamic spatial layout, however this simplistic distinction obscures the complex contributions of each of these distinct pathways to the visual perception of dynamic spatial layout.

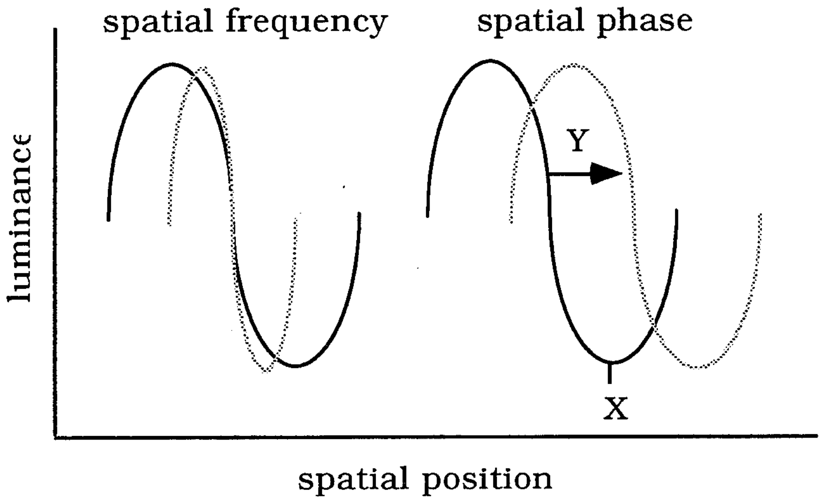

Figure 3. Luminance distribution indicating changes in spatial frequency and spatial phase which are reflected in the spatio-temporal tuning of X and Y cells. While X cells are phase-specific, Y cells demonstrate responding to phase modulation or drift.

Dendritic field sizes and receptive fields of magnocellular cells are significantly larger than those of parvocellular cells. Specifically, magnocellular cells typically exhibit receptive field centers approximately three times larger than their parvocellular counterparts (Lennie, et al., 1990). Among both magnocellular and parvocellular cells, receptive field sizes increase with increasing eccentricity, however magnocellular cells are more frequently encountered in the periphery whereas parvocellular cells are more concentrated foveally. These distinguishing characteristics were first evidenced in physiological research on receptive properties of retinal ganglion cells in the cat (see DeValois & DeValois, 1980 or DeValois & DeValois, 1988 for a review). These distinct classes of retinal ganglion cells in the cat were classified as X- and Y-cells[1]. While X-cells have been demonstrated to be narrowly tuned to certain spatial frequencies and are phase or position specific , Y-cells demonstrate broader tuning to lower spatial frequencies and sensitivity to phase modulation or drift (see Figure 3). The nature of the receptive field characteristics among these distinct classes of cells indicates that Y-cells play a role in the analysis of temporal changes and motion, whereas the X-cells provide precise information regarding location and form.

These X and Y cells project primarily to the lateral geniculate nucleus (LGN). Projections to the superior colliculus are also evident among Y cells. These collicular projections appeared early in the evolution of the visual system and appear to play a major role in motion detection and guidance among lower species. In more elaborately evolved visual systems, the projections to the thalamus predominate and serve to distinguish separate parallel pathways along which these X and Y cells, or P and M cells as they are frequently referred to in primates, travel (cf., Lennie, et al., 1990). These functionally divergent pathways are termed magnocellular and parvocellular in reference to the laminae or layers in the LGN to which these X and Y cells project. The parvocellular layers receive inputs exclusively from X cells, whereas magnocellular layers receive inputs from both X and Y cells, although projections from Y cells predominate. The magnocellular and parvocellular cells can be differentiated primarily on the basis of their differential temporal patterning, and to a lesser extent, in terms of their spatial properties. In particular, magnocellular cells demonstrate greater temporal sensitivity and are broadly tuned to low spatial frequencies whereas parvocellular cells are finely tuned to increasingly higher spatial frequencies (see Figure 4). Further differences among these systems include the sensitivity to color contrast among parvocellular cells that is not evident in magnocellular cells. This suggests that the parvocellular system plays a role in color vision. However, the distinction of these two systems on the basis of color sensitivity is of little interest in the present inquiry since evidence has repeatedly demonstrated dependence on luminance contrast for the computation of spatial and motion percepts (Cavanagh, 1987; Livingstone & Hubel, 1988). Nonetheless, percepts of

[1] Generally, X cells are considered components of the parvocellular system, and Y cells components of the magnocellular system. Although, some X cells do project to magnocellular layers of the LGN.

position, form, and motion reflect a complex symbiosis of the magnocellular and parvocellular systems. However, at early levels of processing these two parallel pathways act independently to analyze information regarding form and motion. The manner in which specific form segregation is computed primarily through the parvocellular system will now be considered in reference to its role in the determination of spatial layout. Subsequently, the role of the magnocellular pathway in early motion detection and the integration of magnocellular and parvocellular inputs in the computational recovery of dynamic spatial layout will be considered in detail.

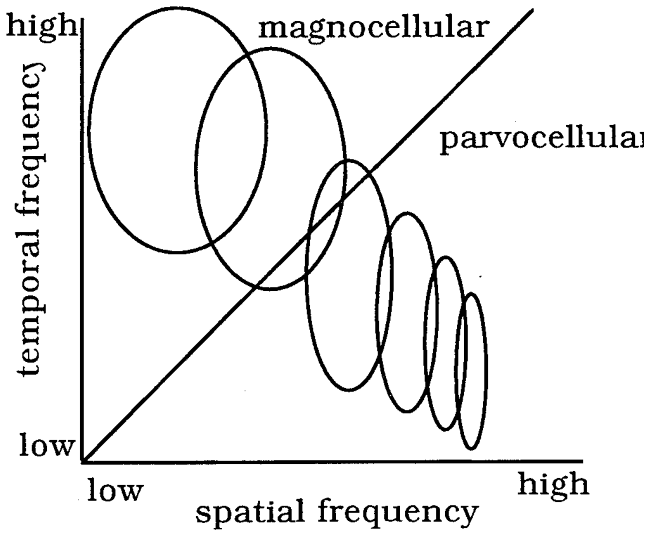

Figure 4. A schematic representation of the division of spatio-temporally tuned filters between the magnocellular and parvocellular pathways of the visual system (contrast with Hess & Snowden, 1992).

Spatial Filtering, Edge Detection, and Form Segregation

Cells of the striate cortex receiving projections from the parvocellular pathway characteristically demonstrate tuning to spatial frequency and orientation. That is, cells demonstrate optimal responding to sinusoidal gratings of particular spatial frequency and orientation placed within their receptive fields (see DeValois & DeValois, 1988). Deviations from this optima are characterized by decreases in cell firing rates that follow an inverted-U function of spatial frequency or orientation. Based upon these response characteristics, it has been hypothesized that these cells act as filters that decompose the luminance profiles of a visual scene into its fundamental spatial frequency components across a series of ranges. Research has suggested that there may be a few as six such spatial frequency filtering mechanisms in human vision, demonstrating peak responding at 0.8, 1.7, 2.8, 4.0, 8.0, and 16.0 cycles per degree of visual angle (Wilson, 1986). Across these channels, spatial frequency bandwidths, measured by the degree of change in spatial frequency that decreases responding to 50% of optimal firing, decrease from approximately 2.5 to 1.3 octaves. The characteristics of cell firings across these channels of progressively higher resolution can provide information about luminance contrast in a local area. Thus, the overall contrast sensitivity function of a perceiver can be seen as an envelope comprised of these more narrowly tuned neural channels (see DeValois & DeValois, 1980; DeValois & DeValois, 1988 for detailed reviews). Pooling from these local orientation and spatial frequency selective channels can consequently serve to provide detailed luminance contrast information to higher levels of the visual system. Tootell, Silverman, and DeValois (1981) examined the organization of spatial frequency selective cells in the primary visual cortex. Findings indicated a regular columnar arrangement of cells as a function of spatial frequency. More recent research by Silverman, et al. (1989) has indicated that there is a definite regularity and systematic organization of cortical neurons that are particularly tuned to various spatial frequencies. Evidence from these studies suggests that there is a regular, periodic anatomical arrangement of spatial frequency tuning among cells in the striate cortex. Furthermore, there is a unimodal distribution of peak tuning with the majority of cells demonstrating the greatest activity in response to intermediate spatial frequencies. From these findings, it appears that the primary visual cortex is comprised of neurons arranged into modules of spatial frequency and orientation selectivity. Based upon this physiological evidence, DeValois and DeValois (1988) have proposed a model of the striate cortex in which their is an orthogonal arrangement of spatial frequency and orientation selective columns. While this model only offers a rudimentary description of feature processing at relatively low levels of cortical processing, it can be posited that the conglomeration of firing patterns within this matrix of orientation and spatial frequency selective columns could produce a filtered spatial frequency "signature" of a given visual scene which can subsequently be integrated at higher levels of visual processing.

David Marr and his colleagues (see especially Marr & Hildreth, 1980; Marr, 1982) offer a means through which this primary spatial filtering can provide information for edge detection. Marr's algorithm posits that visual information is processed in several functional stages. The first stage, which could be physically carried out by low-level spatial frequency selective neurons, serves to filter the visual scene across local regions. This process of filtering can provide inputs to a second stage in which luminance information conveyed through these filtered signals can be differentiated with respect to spatial position. In particular, this later processing would serve to perform a function equivalent to mathematically taking the second derivative of the luminance profile with respect to spatial position or location in the low-level spatial representation. This function would serve to extract what Marr terms *zero-crossings* that refer to edges in the visual scene. Thus, "the model of Marr...postulates spatial filtering as a preliminary stage to a system of edge detectors" (DeValois & DeValois, 1980, p. 330). Marr's (1982; cf., Marr & Hildreth, 1980) theory posits that the visual system progresses through several stages, the first serving to filter two-dimensional luminance contrast information across the visual scene. The spatial frequency filtering characteristics of neurons in the visual system could potentially carry out this requisite filtering during early stages of visual processing. Later cells in the visual system could subsequently serve to extract edges from these filtered signals. According to Marr (1982), this process is accomplished by spatial differentiation of the luminance profile defined through these filtered signals. The actual algorithm employed in Marr's (1982) model involves the convolution of the two-dimensional image array with the Laplacian transform of a Gaussian filter function ($\nabla^2 G * I$). The result of this process is referred to as the *raw primal sketch* and conveys information regarding the spatial location of perceptible contrast changes in the optical array. The position of these contrast changes are termed *zero-crossings* because they refer to points at which the algorithmic second-derivative function has a value of zero. While these *zero-crossings* provide a rudimentary representation of featural and textural boundaries in the visual scene, the contrast changes that they signal may be produced by shadows or noise and therefore are not sufficient indicators for object form segregation. Consequently, several theoretical accounts attempt to describe the subsequent processing of this low-level edge information to perceptually recover form boundaries.

Marr (1982) proposed that locally, orientation and contrast similarities along with proximity can serve to recursively aggregate zero-crossing points into common place tokens defining regions of uniform contrast or texture discontinuities (cf., Julesz, 1981; Bruce & Green, 1985). Essentially, these place tokens group local edge features based upon the similarity of their characteristics such as proximity, orientation, and curvature. This process of feature aggregation can operate through iterative computations of successively higher feature levels and has three principal forms or constraining properties imposed upon the grouping of features. The first of these is position aggregation or clustering whereby edge features in local proximity come to be organized into a perceptual group. Second, curvilinear aggregation allows for the recovery of continuous curved demarcations or boundaries through iterative groupings of local features of similar orientation through the imposition of constraints similar to the Gestalt principles of closure and good continuation (cf., Bruce & Green, 1985). Finally, a process Marr (1982) terms *theta aggregation* allows for the recovery of textural features or illusory contours as they are sometimes termed through the grouping of similarly oriented textural details. The specific algorithmic techniques for recovering these visual features is beyond the scope of the current inquiry, but its general strategy is notable in that it reflects the implementation of higher level or top-down constraints on object characteristics that lead to the recovery of feature boundaries. This notion of processing constraints is significant and will be reintroduced several times throughout this inquiry regarding visual spatialization in reference to high level recovery of monocular and binocular motion and position information. Currently, discussion will focus on these top-down constraints in reference to a recent theoretical account of form segregation: Biederman's recognition by components (RBC) model of feature detection.

Biederman's (1987) model of object component recognition is particularly appealing as a description of the segregation of form boundaries because it focuses on a small set of unique features from which all conceivable two-dimensional views of three-dimensional objects can be constructed. Therefore, this unique set of features constrains the perceptual recovery of object boundaries by indicating those edge groupings that can be considered components of a solid three-dimensional geometric object in the visual scene. Biederman (1987) suggests that a relatively small set (about 36) of geometric generalized-cone components termed *geons* comprise the set of sufficient features for object construction and these *geons* can be perceptually recovered through determination of two-dimensional edge properties of curvature, collinearity, symmetry, parallelism, and codetermination. Indeed, this theory describes a precise means through which the iterative computational aggregation of local edge information can lead to the recovery of object boundaries by the imposition of higher level constraints regarding the structure of featural components. Indeed, Biederman's RBC model can be seen as an extension of Marr's own theoretical analysis of object recognition in which he and his colleagues similarly suggest that generalized-cone components are the constraining features of three-dimensional objects and can be perceptually recovered through analysis of the two-dimensional arrangement of object surface characteristics (see Marr & Nishihara, 1978; Marr, 1982). Thus, current theoretical approaches to form segregation posit that the demarcation of object boundaries in the visual scene can be accomplished through a series of computational processes involving the initial decomposition and filtering of the optic array, the subsequent recovery of edges or contrast boundaries, and finally the perceptually grouping of these boundaries through the imposition of constraints regarding the grouping of local edge features.

This cursory review of the vast literature on the processing of luminance contrast information and the perception of form in the visual system has only touched upon the major themes represented in these writings. Further inquiry into the nature of form processing is clearly critical for a complete analysis of the dynamic spatial layout and consequently the analysis of visual form will be reintroduced at various points throughout this consideration of visual spatialization in reference to the computational recovery of high-level object motion information, the segregation of texture and object features that cue depth, and the determination of feature correspondence in the recovery of binocular disparity information. Presently, the consideration of visual spatial processing will focus upon a detailed consideration of the perceptual recovery of motion information. In this analysis of motion, it is posited that the rapid detection of motion, mediated primarily through the magnocellular pathway, is subsequently augmented at higher levels by texture and feature information provided predominantly by parvocellular analyses of luminance contrast through computational processes such as those previously described.

Motion Detection

Based upon the functional distinction between transient magnocellular and sustained parvocellular responding, these systems have been relegated to tasks of motion detection and form identification, respectively. However, a complete representation of the dynamic spatial layout requires the integration of information regarding motion and form. A theoretical consideration of how this motion and form information is integrated to create a dynamic percept of spatial layout follows. This integrated model suggests that the need for rapid information regarding motion and precise information regarding form and position is achieved through two processes, one for rapidly extracting dynamic motion information, the other for providing a more complete account of the structures and patterns of the dynamic spatial layout.

The notion of these two distinct processes for visual motion detection was introduced by Braddick (1974, 1980). The first of these processes, termed the "short-range" process, was presumed to operate under conditions of visual presentation that occur over a small spatial extent and short exposure duration. Consequently, such a process is characteristic of properties typically attributed to the magnocellular pathway. The motion solution produced by this "short-range" process is arrived at relatively early in the stream of visual information processing. On the contrary, the second process, termed the "long-range" process, operates on stimulus motion over greater spatial extents and longer durations and occurs relatively late in visual processing. Furthermore, the "long-range" motion process is assumed to be partially mediated by form information as well as top-down constraints on object form and motion. Consequently, this "long-range" motion solution can be seen as an integration of magnocellular and parvocellular inputs into a recombined solution reflecting object form and motion. This presence of two distinct processes for the detection of motion can serve to greatly enhance the dynamic range over which the visual system is sensitive to motion in the visual field. Moreover, this architecture is particularly adapted to the rapid detection of motion as well as subsequent analyses of visual form.

Braddick (1974) originally proposed this two-process distinction based on results of experiments using random dot kinematograms. Random dot kinematograms can be constructed by initially assigning pixels in a bitmap array to different luminance contrast values. Subsequently, in a series of display frames, a local region in this random array can be displaced in a systematic fashion. Other pixel elements in the display may either remain stationary or may be displaced in an uncorrelated fashion. Thus, the parameters that can be manipulated in such displays include the absolute displacement of the region between successive frames, the interval between these successive frames (the interstimulus interval or ISI), the luminance contrast between pixels, and the color contrast between pixels. When these parameters are set within a limited range, the local region of orderly displaced elements is perceptually segregated from its background and is perceived as a uniform object in motion. The specification of these delimiting parameter values serves to delineate the boundaries within which the "short-range" low-level motion system operates.

In reference to constraints placed upon the uniform displacement of elements in the display, minimal and maximal values for the veridical perception of motion direction can be obtained. These values, D_{min} and D_{max}, have been obtained using various stimuli including random dot kinematograms and sine wave gratings (see Sekuler, et al, 1990; Boulton & Hess, 1990a & 1990b). Within the range over which motion is veridically perceived, psychometric functions of accuracy in the estimation of motion direction exhibit an inverted-U distribution with peak sensitivity to motion at an intermediate displacement value termed D_{opt}. Amongst these delimiting values, D_{max} is particularly important because it specifies the boundary between conditions under which the "short-range" and "long-range" processes operate. That is, when displacement is less than D_{max} movement can be perceived in a sequence of display frames without a concomitant determination of object form, whereas displacement over distances greater than D_{max} requires further assumptions regarding the consistent form of the displaced object. Braddick's early work (see Braddick, 1974), indicated that the value of D_{max} was a constant assuming a value of approximately 15 arc minutes of visual angle. More recent research has indicated that this original conclusion was erroneous and D_{max} increases linearly with retinal eccentricity with displacements of up to 90 arc minutes being perceived coherently beyond 10^{o} of eccentricity in the peripheral field (see Sekuler, et al., 1990). Moreover, D_{max} is highly dependent on spatial frequency (Anstis, 1986; Sekuler, et al., 1990; Watson, 1990; Boulton & Hess, 1990a & 1990b). Recent research has demonstrated this interaction between maximal displacement and spatial frequency using sequential frame presentations of sinusoidal gratings under conditions of varied phase displacements between frames. By representing D_{max} in relation to spatial frequency, the maximal displacement assumes a value between 1/6 and 1/2 of the spatial wavelength depending on what particular study is cited (see Sekuler, et al., 1990; Boulton & Hess, 1990a & 1990b). Indeed, motion can be most economically and efficiently conveyed through quadrature pairs of motion detection units having receptive fields spaced at 1/4 wavelength intervals (see Adelson & Bergen, 1985). Consequently, it is predicted that a displacement of a sinusoidal grating of 1/4 wavelength will be optimal with respect to the discrimination of the direction of apparent motion. However, psychophysical studies indicate that these spatial motion detection units might be separated by less than 1/4 wavelength, since D_{opt} exhibits values of approximately 1/5 or 1/6 of the spatial wavelength (Boulton & Hess, 1990a and b). In any case, the perception of "short-range" motion is spatially limited by an upper bound that is interactively determined by the displacement distance and spatial frequency characteristics of the display elements. Moreover, this spatial limitation is dependent upon stimulus contrast. In particular, D_{min} values (but not D_{max} values) demonstrate decrements in sensitivity to motion with increasing contrast (Boulton & Hess, 1990a). Thus, apparent motion over short spatial extents can occur independent of form identification but is limited by a complex interaction between the extent of the displacement and the spatial frequency and contrast characteristics of the stimuli.

The perception of apparent motion in such displays is further limited by temporal characteristics of the display sequence. In general, the perception of coherent motion in random dot kinematograms is diminished with increasing ISI durations and has an upper bound of approximately 40-80 msec beyond which motion cannot be perceived in successive frame displacements (Braddick, 1974; Anstis, 1986; Petersik, 1989). Furthermore, spatial displacement parameters interact with ISI. In particular, D_{max} decreases with increasing ISI (Petersik, 1989). This evidence suggests that the perception of motion is the function of processing a spatio-temporal sampling of the luminance profile across successive frames or scenes. Several important theoretical assertions can be made from this conclusion regarding the spatio-temporal nature of visual encoding. The constraints placed upon the perception of motion by these above discussed spatial and temporal parameters suggests that there is a "window of visibility" delimiting the spatio-temporal range where coherent motion can be perceived by the "short-range" processing of the magnocellular system (cf., Watson, Ahumada, & Farrell, 1986). Consequently, theoretical accounts of how motion information is recovered within this "short-range" system typically involve the differentiation of spatial luminance characteristics with respect to time (see especially, Marr, 1982). The tuning of "short-range" motion detection units is such that within the boundaries specified above, motion can be readily perceived without reliance on a concomitant solution of object form. Recent theories have stressed that low-level motion information is represented as a distribution of spatio-temporal energy derived from the time sampled luminance profile (Watson & Ahumada, 1985; Adelson & Bergen, 1985). These energy models equate the detection of motion with the detection of orientation in a two dimensional scene except that the image vector varies across time instead of space. Such a system could be physically realized by a lattice of phase modulation sensitive cells such as those characteristic of the magnocellular system (see Figure 5). An arrangement of these local phase modulation detectors separated by 1/4 of their tuned spatial wavelengths could signal movement in a particular direction[2]. This could be accomplished through neural temporal encoding in which time sampled energy is compared across motion sensitive cells. Orientation and direction of travel can be signaled by embedded AND and NOT gates in the neural motion detection lattice. This lattice structure can be replicated across the proximal space and demonstrate local cooperativity to signal the direction of motion in regions of the optic array. Consequently, a rapid determination of motion can be determined through spatio-temporal energy changes in the lattice of motion detecting units independent of an analysis of object form.

However, form is readily perceived in random dot kinematograms and distinct edges arise from the apparent motion of a local region. Therefore, it appears that in these instances form information is deduced from lower level motion information as well as higher order constraints. Such a process could be easily defined if it were not for the indeterminacy of correspondences between spatial elements sampled across time, the so called *correspondence problem* . Given the random pixel mapping defining the bitmap of a random dot kinematogram, the correspondence between moving pixels in the display cannot simply be derived at a feature level that operates on individual pixel or photoreceptor elements because the mapping of such an element in one time-sampled image to the correct corresponding element in a second time-sampled image is indeterminate. Yet, coherent motion of a uniform object or region can be extracted from successive displacements across frames suggesting that the correspondence problem is solved at a feature level higher than individual photoreceptor elements (see Marr, 1982; Adelson & Bergen,

[2] While Adelson & Bergen's (1985) spatio-temporal energy model explicitly posits a lattice structure comprised of quadrature paired neurons tuned to optimal phase separations of 1/4 wavelength this merely represents the most economical structure. Therefore, this lattice can be constructed with neurons whose tuning characteristics are more narrowly separated in reference to phase. Therefore, this model in its general form is not discrepant with psychophysical data indicating narrower tuning to spatial phase modulation (see Watson, 1990; Boulton & Hess, 1990 a and b).

1985). This suggests that motion and form information interact at various levels along the visual pathway despite their early segregation. In this manner a uniform percept of object motion and object form can be derived through a successive build-up of information along these pathways in conjunction with input from higher level constraint parameters. Therefore, the level of depth in visual processing at which the "short-range" process occurs is of considerable interest and will now be considered in detail.

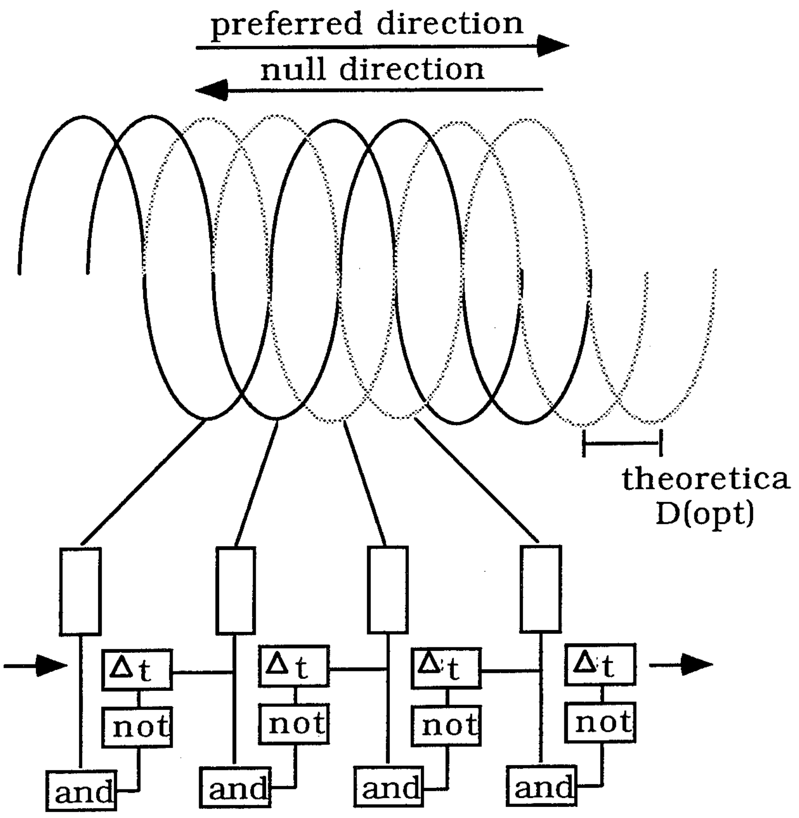

Figure 5. A lattice of quadrature paired local phase modulation detectors that signal motion in a particular direction (after Marr, 1982 and Adelson and Bergen, 1985).

Braddick's (1974) initial investigation of "short-range" motion percepts included conditions in which successive frames of the random dot kinematogram was presented dichoptically, alternating successive frames between the two eyes. Results demonstrated that no apparent motion of the regional displacement was witnessed under these conditions. Similar effects have been well documented and lead to the conclusion that the process of "short-range" motion detection is a peripheral one occurring prior to binocular pooling (Petersik, 1989). Other evidence suggests that the "short-range" detection of motion occurs quite early in the stream of visual information processing. For example, studies conducted using stimuli which alternated stimulus contrast between successive frames resulted in no perceived apparent motion (Petersik, 1989). These results lead to the conclusion that information available to the "short-range" process is not phase insensitive and further suggests that the "short-range" process does not involve a complete solution to the correspondence problem. It has also been demonstrated that the detection of motion at this level occurs independent of chromatic information. Using isoluminant pixels with different chroma values in random dot kinematograms, Anstis demonstrated that these color differences alone were insufficient for producing a percept of motion (see Anstis, 1986). From these results, it can be concluded that the derivation of motion information performed by the "short-range" process occurs early in the visual stream and operates using monocular luminance contrast information. The characteristics of the "short-range" process are strikingly similar to the receptive field characteristics of magnocellular Y-cells, and not surprisingly, the magnocellular system described above is believed to be primarily responsible for this rapid early detection of motion information (see Livingstone & Hubel, 1988; Lennie, et al., 1990). However, this system is not particularly adapted for detecting luminance contrast changes over time. Such luminance contrast changes are frequently encountered when objects traverse a cluttered environment and reflect varying amounts of light due to the characteristics of a non-uniform light source and shadowing effects caused by other environmental features. Therefore, the correct interpretation of moving forms in the visual array requires computational correspondence solutions that are insensitive to changing contrast between corresponding points. Therefore, such a solution is presumed to function at a higher level and involves the pooling of form and motion information. This higher level or "long-range" process is critical for perceiving dynamic spatial layout in a coherent fashion and will consequently be considered in detail.

Perception of form from random dot kinematograms is presumably constrained by the spatio-temporal envelope under which "short-range" motion processing occurs because such displays lack any information in which form can be distinguished independent of motion. In more typical configural displays or environmental scenes, the interpretation of form can have a dramatic influence on the motion percept witnessed by observers. The phenomenological study of such percepts has a rich history associated with the Gestalt movement in psychology during the first half of this century (see, e.g., Bruce & Green, 1985 for a review). One class of stimuli demonstrative of the influences of higher level interpretations of form on percepts of motion is the Ternus display, named after the Gestalt researcher who first performed research using these stimuli (see Marr, 1982; Anstis, 1986; Petersik, 1989). A typical configuration of a Ternus display is shown in Figure 6. In such a configuration, the outer element or elements of a linear arrangement are displaced about a series of inner elements that maintain their position across frames. The outer element or elements are subsequently returned to their initial position and this sequence can be iterated any number of times. In such displays, either the outer elements can be perceived to move to either side of the stationary inner elements, or all display elements can be perceived to undergo a coordinated displacement to and fro even through the two inner elements remain stationary. Research has demonstrated that the erroneous percept of group motion could be abolished by having subjects focus on the central elements (see Anstis, 1986). Subsequent research has demonstrated that under conditions favoring "short-range" processing, the motion is perceived as displacement of the outer elements. However, when conditions of presentation permitted processing at higher levels (i.e., an imposition of grouping strategies) the erroneous perception of group movement predominates (see Anstis, 1986). In particular, biasing perceptions with the Ternus display can be achieved by varying the ISI. With brief ISIs between displacements (less than approximately 40-100 msec), conditions favor the perception of outer element movement, whereas longer ISIs bias perception toward coherent group motion.

Thus, the Ternus effect of perceived group motion can be seen as a function of high level rectification of the correspondence problem through the imposition of constraints. In particular, the Ternus effect provides a condition of indeterminacy with respect to the correspondence problem. Since each element is identical and consistently equidistant across frames, a conclusive correspondence cannot be deduced. It appears that in these instances the observer employs two key heuristics for solving the correspondence problem resulting in the percept of group motion. First, the observer may simply use a proximal judgment heuristic suggesting that correspondence mappings should be made between objects that have displaced the smallest distance from one frame to the next. Second, monocular depth cues may play an important role in this correspondence judgment. Since the objects in the Ternus display appear to be on the same depth plane, a correspondence solution that correctly maps the outer elements to successive positions on either side of the inner elements seems physically impossible because the two inner elements block this motion path.

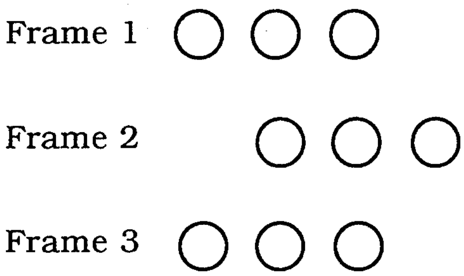

Figure 6. A sample of a typical Ternus display. The outer element is displaced to either side of the central element and these frames can be displayed in an iterative fashion.

Sekuler, et al. (1990) describes evidence that several other heuristic assumptions of this type influence motion processing at higher levels. The first of these assumptions is that of *inertia*, which states that objects in motion will tend to stay in motion. The impact of this assumption can be seen in displays of rotary motion. If only two frames are presented in a rotary motion display, the direction of motion is ambiguous and can be perceived as being either clockwise or counterclockwise. However, if a third frame is presented, the sequencing of these three frames is sufficient to produce a unambiguous percept of a single direction. The second assumption is that of *rigidity* which holds that all features of a moving body move approximately in synchrony. This assumption may allow for partial solutions to the correspondence problem based on more global structural features, but is often violated in dynamic real world scenes. For example, Gunnar Johannson and his colleagues (see Johannson, von Hofsten, & Jansson for a review; cf., Bruce & Green, 1985) have extensively studied the perception of biological motion in which global rigidity is absent. However, in the motion of biological organisms, rigidity does exist at local levels. That is, movement occurs only at the joints which serve to demarcate regions of local rigidity. Research has demonstrated that presentation of the synchronized movements of just these joint locations, accomplished by placing lights on the joint areas of a moving organism in an otherwise completely darkened viewing environment, leads to high rates of correct object identification. In essence, it can be concluded that the recovery of form from the motion of these discrete points is accomplished by a perceptual filling-in of regions of rigidity that lie between the lighted joints. The perceptual recovery of form information is believed to be the product of a vector analysis of the discrete points that provides information regarding their relative movement which consequently places constraints upon grouping strategies. Therefore, an assumption of rigidity at local levels using relative motion information allows for the build-up of a global form representation. A third assumption, closely related to the concept of rigidity, is that of covering and uncovering or the perceived *persistence* of a object with occlusion (Sekuler, et al., 1990). For example, in a two frame presentation in which two objects are presented in the first frame and only one of these objects is presented in the second frame and a simple path of motion can be assumed, a percept that the missing object in the second frame has moved behind the other object can be formed. Thus, "...the visual system seems to assume that an object continues to exist even if the system has to fabricate the supporting evidence" (Sekuler, et al., 1990, p. 219). Movement in a cluttered environment is particularly prone to occlusion during the course of travel. Consequently, an interaction of lower-level and higher level processes must occur to extrapolate the rate and direction of movement of a particular object that becomes momentarily hidden from view. First, an occluded object is presumed to maintain form and maintain trajectory characteristics demonstrated prior to occlusion (see Hochberg, 1986). Second, the constant velocity or acceleration of the moving object can be accurately estimated with increasing exposure prior to occlusion thereby allowing relatively precise estimation of location during occlusion (Rosenbaum, 1975; Jagacinski, Johnson, & Miller, 1983). Consequently, it is evident that higher-level constraints are placed upon the motion of occluded objects that allows for persistent representation of their spatio-temporal characteristics through estimated "cognitive trajectory" parameters (Jagacinski, Johnson, & Miller, 1983). The quality of these "cognitive trajectory" representations diminishes over time due to noise and equivocation of information regarding the objects spatio-temporal properties and consequently spatial judgments become increasingly variable with increasing time in occlusion. Nonetheless, the ability to make estimates regarding the position of hidden moving objects demonstrates the importance of top-down constraints such as *inertia*, *rigidity*, and *object permanence* in the construction of a spatio-temporal representation of the dynamic spatial layout of the environment.

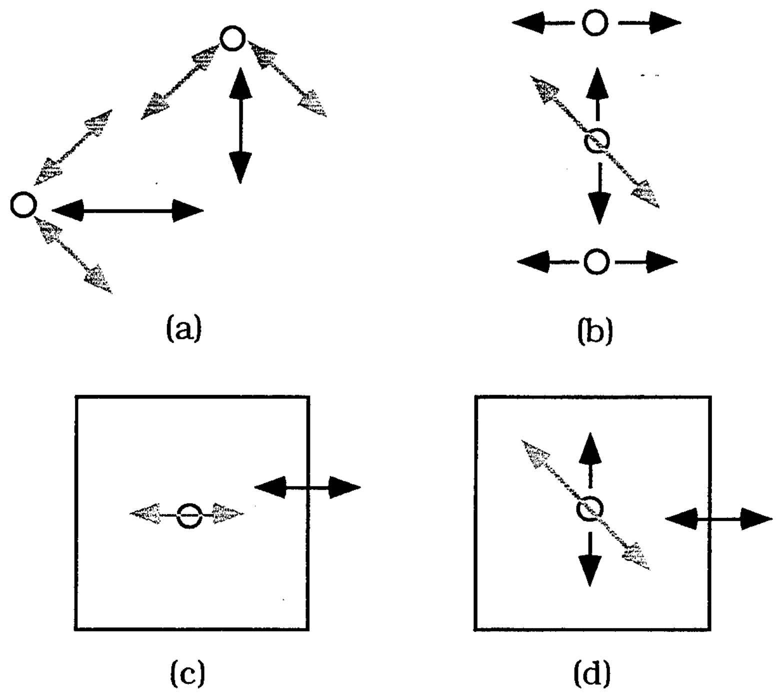

Figure 7. Examples of relative motion trajectories (gray lines) that deviate from the veridical motion of the objects (black lines) and are induced by the motion of neighboring or adjacent objects (from Gogel, 1974 and Gogel, 1978).

Thus, it appears that a higher level segregation of form information leads to the implementation of constraints regarding object properties and is subsequently combined with lower level motion information to produce an overall percept of spatial dynamics. As demonstrated in the Ternus display, boundary conditions arise when form information is ambiguous and attempts to impose figural constraints in these scenarios results in non-veridical percepts. These boundary conditions between "short-range" and "long-range" motion are further demonstrated in the *adjacency principle* proposed by Walter Gogel (1974; 1978; cf., Johansson, von Hofsten, & Jansson, 1980; Mack, 1986). In the Ternus display, the impact of proximity constraints appear to be a major contributing cue to the percept of group motion. Gogel (1974) has demonstrated that in other configural displays, the perceptual grouping of two distinct features in motion is a function of their adjacency or fronto-parallel separation. In particular the *adjacency principle* states that "...the effectiveness of cues between objects is an inverse function of object separation" (Gogel, 1974, p. 425). Consequently, when two adjacent features are placed in motion along orthogonal axes in which their absolute motion vectors are separated by 90^{o} (black lines in Figure 7a), perceptual grouping of these items leads to a motion percept that tends toward their common motion vectors (gray lines in Figure 7a; Gogel, 1974; Gogel; 1978; Johansson, von Hofsten & Jansson, 1980). With decreasing separation between items, the perceived motion vector approaches the common motion vector and is less influenced by the veridical relative motion vectors of the objects. On the contrary, when an object is imbedded within another object, the objects are perceptually segregated into figure and ground. Therefore, the assumption of moving objects on a stationary ground leads to the percept of inner element motion even when the veridical percept is one of frame motion (Figure 7c). Similarly, when the inner element and the frame move along orthogonal axes, the frame is perceived to remain stationary while the inner element is perceived to move along a relative motion vector (Figure 7d). Figure 7b demonstrates a particularly difficult case. In this scenario, the outer elements can either be segregated from the inner element or be grouped together with it. Typically, segregation occurs through implementation of a simplicity constraint and the percept that arises is one of inner element motion along a relative motion vector and stationary outer elements. That is, if one assumes a grouping of the items, a motion-in-depth or changing size percept would arise. However, no other cues are available to substantiate a motion in depth percept and changing size in inherently complex and unlikely. Therefore, the simple relative percept of inner element motion along a relative motion trajectory arises.

In all of these cases, the percept of motion is a relative one. That is, the perceived direction of motion is dependent upon grouping strategies and the relative motion vectors of objects as compared to adjacent features. As separation between display items increases, motion becomes more veridical with respect to the absolute motion vector. Therefore, high-level form processing and the imposition of featural constraints can be presumed to mediate percepts of relative and common motion. Of course, the magnocellular system's low spatial resolution renders it poorly adapted for the segregation of multiple features in motion over a small spatial extent. Therefore, the high-resolution parvocellular system apparently mediates this phenomenon by providing partial form solutions which are subsequently subjected to high level featural constraints. When separation becomes sufficient, absolute motion predominates for two reasons. First, the distance can be sufficient for discrete analysis by low spatial frequency tuned motion detectors. Second, the large separation renders higher resolution form detectors to conclude that the features are components of distinct objects. Thus, over short spatial extents, the adjacency principle posits an intermediating role of "long-range" motion detection to give rise to movement along common or relative trajectories, whereas over larger spatial extents, "short-range" and "long-range" motion cues signal motion along absolute trajectories in a fashion consistent with the spatio-temporal tuning characteristics of magnocellular and parvocellular input pathways. Thus, as evidenced in the Ternus display and the relative motion trajectories of adjacent objects, the "long-range" motion detection process appears to place influential top-down figural constraints on the percept of spatial dynamics.

Consequently, the "long-range" motion process described thus far appears to be influenced by an analysis of object form in addition to top-down constraints imposed by a consideration of this form information. Therefore, it is evident that such solutions occur later in the visual stream than solutions attributed to the "short-range" process. Psychophysical evidence supports this conclusion of a late motion processing stage reliant upon form processing and the imposition of top-down constraints. First, dichoptic presentation of configural motion displays such as the Ternus display does not diminish the perception of motion (Petersik, 1989). This suggests that the "long range" process operates on pooled binocular information. Furthermore, apparent movement of isoluminant, color contrast stimuli is perceptible to the "long-range" process, whereas the "short-range" process is sensitive only to luminance. Finally, changes in contrast and even changes in form characteristics of objects across frames does not destroy perceived object motion, suggesting that the "long-range" process does not operate at the level of photoreceptor elements, but rather it functions at some higher level of abstraction of that information. In fact, this "long-range" process appears to be the result of a representation of spatial layout and spatial dynamics that arises from a pooled interaction of magnocellular and parvocellular inputs.

Many of the key findings regarding the "short-range" and "long-range" processes have been described above. These findings have been replicated using stimuli other than the random dot kinematograms for "short-range" effects and the Ternus display for "long-range" effects, but these stimuli are those most frequently associated with this line of research. "Short-range" phenomena have also been studied using sinusoidal gratings and other narrow-band stimuli, while "long-range" processes have been examined with a myriad of configural displays which are sometimes filtered to examine spatial frequency effects of such stimuli. The general conclusions of these research efforts are summarized in Table 1.

Table 1. The principal findings regarding "short-range" and "long-range" motion processes (sources: Marr, 1982; Petersik, 1989).

Short-Range	Long-Range
Operates over small spatial extents - approx. 15' centrally up to 90' in the far periphery	Operates over a large spatial range - several degrees
Requires small ISIs - <40-100 msec	Preference for long ISIs - 300 msec or greater
Bright ISI destroys motion perception	ISI may be bright or dark
Motion effect cannot be produced dichoptically	Dichoptic presentation results in motion percept
Requires luminance contrast gradients	Chroma differences between stimuli and background is sufficient
Sensitive to luminance and contrast changes across frames	Not sensitive to luminance changes
Sensitive to form changes across frames	Not affected by form changes in the stimuli
Motion detection precedes edge extraction	Edge extraction precedes motion detection
Prefers low spatial frequencies	Prefers high spatial frequencies

From this evidence a reasonable assumption can be made that the "short-range" motion solution functions as a precursory peripheral process that provides information to the higher level, more central "long-range" process. Further evidence for this conclusion is provided by studies of motion aftereffects (see Anstis, 1986; Petersik, 1989). Motion aftereffects appear to be the function of the low-level "short-range" process and in particular, they can be attributed to fatiguing of low level directionally selective receptors. Indeed, motion aftereffects are limited in their spatio-temporal extent in a manner consistent with the aforementioned properties of the "short-range" process (Anstis, 1986). However, motion aftereffects exhibit interocular transfer and the degree of transfer is highly correlated with measures of binocular depth thought to be the function of more central processes (Anstis, 1986). This seems to defy the conclusion that motion aftereffects can be sufficiently accounted for by the "short-range" process. Rather this finding leads to the logical conclusion that the higher level "long-range" process serves to integrate low level motion information from the "short-range" process with other information from the visual display such as information about depth and object form. Much of the other research previously discussed appears to converge on a similar conclusion and allows for the conceptualization of a generalized framework for understanding the interactions between the "short-range" and "long-range" motion processes. In the following discussion, models that specify how this interaction might be computationally specified will be considered in detail.

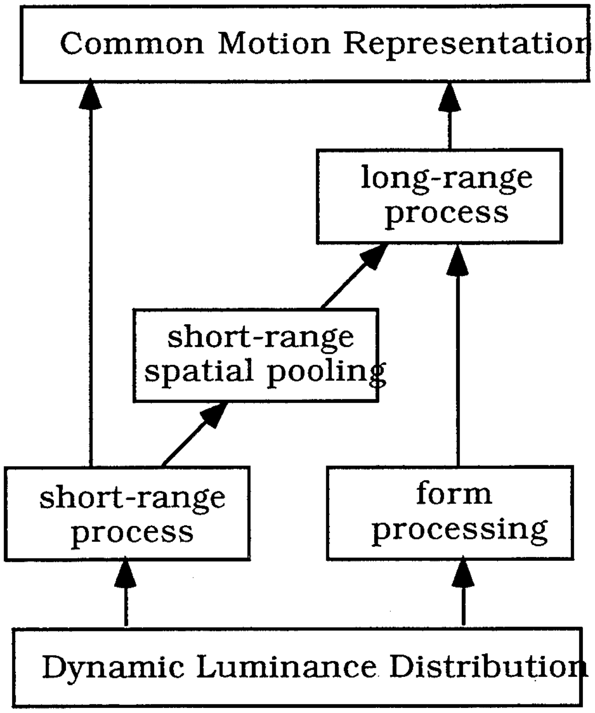
Common Motion Representation
long-range process
short-range spatial pooling
short-range process
form processing
Dynamic Luminance Distribution

Figure 8. Petersik's framework for describing the relationship between short-range and long-range processes with respect to a single unified motion percept. (from Petersik, 1989, p.124).

Based upon physiological and psychophysical data distinguishing the two parallel pathways and two motion detection processes discussed above, a theoretical models of visual spatial processing can be formulated. Petersik (1989) has developed at least a partial account of this processing which is replicated in Figure 8. However, this framework is incomplete because it obscures many of the subtle distinctions between the "short-range" and "long-range" motion and their relationship to parallel pathways in the visual system. First, what Petersik (1989) refers to a "Form Processing" encompasses several distinct functions such as luminance segregation, edge detection, and structural processing as well as color, depth and binocular analysis of the optical array all of which are mediated predominantly by the parvocellular system. Moreover, the rapid detection of spatio-temporal change giving rise to "short-range" motion solutions is largely a function of the magnocellular system. Consequently, the subsequent "long-range" motion solution appears to be the result of some computational weighting processes that operates on the pooled information from the magnocellular and parvocellular system. As demonstrated by interocular transfer of motion aftereffects discussed above (cf., Anstis, 1986), the input of "short-range" magnocellular motion signals to higher levels accrues with increasing exposure durations. Consequently, inputs from the "short-range" process appear to have both direct and indirect or mediated inputs to a common representation of motion in the optical array that reflects the spatial dynamics of the environment. That is, the rapid "short-range" solution functions interactively with higher level feature information and imposed motion constraints to yield a "long-range" motion solution. The overall dynamic spatial representation can consequently be seen as the result of weighted additive and interactive inputs from motion detectors, feature detectors, and top-down form and motion constraints. Moreover, this common spatio-temporal representation can be incrementally built-up and revised in a cascading fashion. That is, the common dynamic spatial or motion representation formulated initially based upon rapid low level "short-range" inputs need not be entirely reformulated in the face of discrepant information from higher level "long-range" processes but rather can simply be re-parameterized to reflect the weighted inputs of form and motion integration accrued over time. In this manner, the visual system can provide rapid motion information and subsequent structural information upon which a spatio-temporal representation of the dynamic optic array can be incrementally constructed over time. This structure has been well adapted to the critical detection of motion from an ecological standpoint. As Marr (1982) notes:

> Perhaps the key to the puzzle is that in the analysis of motion - more so, perhaps, than any other aspect of vision - time is of the

> essence. This is not only because moving things can be harmful, but also because, like yesterday's weather forecast, old descriptions of the state of a moving body soon become useless. On the other hand, the detail of the analysis that can be performed depends upon the richness of the information on which the analysis is based, and this in turn is bound to depend upon the length of time that is available to collect the information (Marr, 1982, p. 162).

The evidence that there are two distinct stages in the stream of visual processing is significant. Moreover, it highlights the interaction of the magnocellular and parvocellular pathways in constructing a unique representation of the dynamic spatial layout that conveys both form and motion information. Up to this point, the recovery of form and motion information has only been considered in reference to the optical structure along the fronto-parallel plane. However, the visual environment is inherently three-dimensional in character and consequently the representation of spatial dynamics and spatial layout must include reference to position and changes in position along the depth axis. Therefore, discussion will now turn toward a detailed consideration of the computation of motion-in-depth.

Dynamic Cues for Depth Perception

The invariant dynamic cues to depth conveyed through local and global flow in the ambient optic array was an area of particular interest in Gibson's (see e.g., Gibson, 1979) ecological approach to visual perception. In Chapter 1, it was demonstrated how these flow patterns produce invariant optical information regarding the motion in depth of environmental objects and the perception of self motion. At this juncture, the neural analysis of these invariant patterns in the flow of the dynamic optical array and subsequent synthesis into a unique spatio-temporal representation of the visual environment will be described in detail. As Gibson (1979) demonstrated, the rate of movement in depth of an object or of the perceiver is specified by the rate of global or local optical flow. David Lee and his colleagues (see Lee, 1980 for an overview) have demonstrated the importance of this optical flow rate in guiding motor actions. David Lee (1980) has described the invariant properties of this optical flow rate as giving rise to information regarding the time-to-contact with the viewpoint of objects in motion relative to the perceiver. Psychophysical and physiological research conducted by David Regan and his colleagues (see

Regan, Beverley & Cynader, 1979 for an overview) has demonstrated a biological basis for the computational processing of such motion-in-depth information within the visual system. Therefore, inquiry into these dynamic cues for the perception of depth will begin with a precise description of the optical flow rate and will subsequently turn toward a detailed analysis of how this information may come to be represented through neural processing.

Rectilinear motion of an object toward the perceiver or motion on behalf of the perceiver produce local or global expansion of the optical array along radial flow lines emanating from the point of expansion (cf., Gibson, 1979; see Figure 2). This pattern of expansion provides unique invariant information conveying the rate of motion in depth and the time-to-contact of environmental objects with the viewpoint at any given time (Lee, 1980; see Figure 9). In Figure 9, the manner in which the rate of motion in depth is uniquely specified through the expansion of proximal image size is illustrated. The image of a distal object, R, at a given time, t, is cast upon the perceiver's retina forming a time-sampled proximal image, r(t). The rate of motion toward the viewpoint or focal point of the lens, V(t) is directly proportional to the rate of expansion of the proximal image, v(t). Consequently, the time-to-contact of the distal stimulus with the viewpoint can be determined through the following series of spatio-temporal computations. Since the distance from the nodal point of the lens to the retina is constant it is assigned a value of unity for computational simplicity. Therefore, based upon the geometry of similar triangles it can be shown that:

$$\frac{Z(t)}{R} = \frac{1}{r(t)} \tag{1}$$

By differentiation of these size distance relationships with respect to time, relative velocity in depth can be specified as follows:

$$\frac{R}{V}=\frac{r(t)^2}{v(t)}$$

where

$$V=\frac{-dZ(t)}{dt} \text{ and } v(t)=\frac{d\,r(t)}{dt} \tag{2}$$

since

$$R=Z(t)r(t),$$

$$\frac{r(t)}{v(t)}=\frac{Z(t)}{V}=\tau$$

This derived higher-order optical variable specifying the relative velocity of approaching environmental objects, termed tau (τ), is critical for specifying parameters for the timing of responsive motor actions relative to objects in the environment (cf., Lee, 1980). When velocity is constant, tau(τ) uniquely specifies time-to-contact of the object with the viewpoint. However, when an object exhibits acceleration, tau (τ) will overestimate the time-to-contact (Lee, et al., 1983). Therefore, estimations of contact times of accelerating objects must include a "delay parameter" that compensates for the change in relative velocity since the last instance of sampling the dynamic optic array. Research has demonstrated that tau (τ) is a sufficient optical variable for specifying timing information regarding the movement of objects in depth and accounts for response performance in estimating arrival times of various environmental stimuli (see, e.g., Todd, 1981; Schiff & Oldak, 1990; Caird & Hancock, 1992). This estimation of arrival time has been demonstrated be a crucial factor for guiding motor actions directed at objects moving in depth in tasks such as hitting and catching a ball (see Todd, 1981; Lee, et al., 1983; Boostsma & Peper, 1992). Additionally, the specification of higher order relative dilation and minification of features in the optic array uniquely specified by tau (τ) have been demonstrated to play a critical role in the maintenance of posture and the negotiation of obstacles during egomotion and the maneuvering of transportation vehicles (Bruce & Green, 1985; Caird & Hancock, 1992). Consequently, rectilinear motion-in-depth comes to be represented by the spatio-temporal information conveyed in the relative rate of expansion of proximal images and provides critical information utilized in subsequent motor planning and the programming of responses directed at environmental objects with three-dimensional dynamic properties.

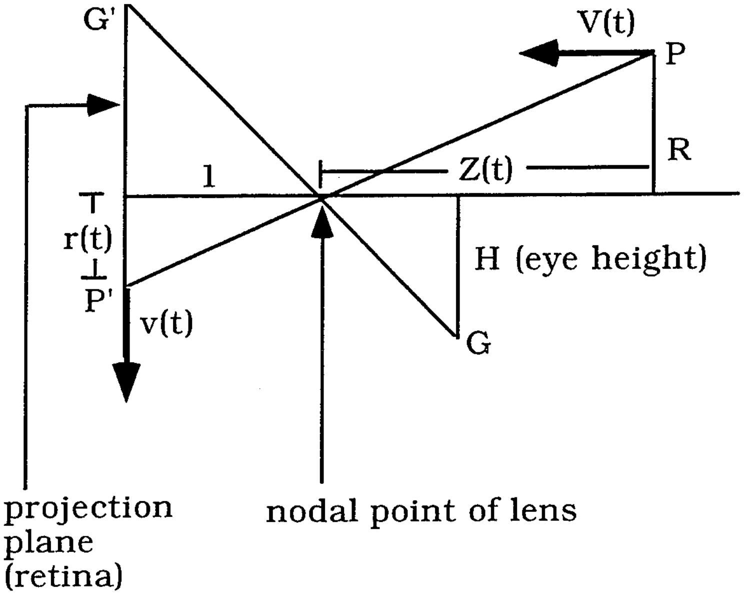

Figure 9. Representation of the optical specification of the relative velocity of an object or self motion (V(t)) conveyed through expansion of the proximal size of environmental features (r(t)). (after Lee, 1980)

Earlier, the neural specification of motion along the fronto-parallel plane was discussed in reference to two processes: (1) the "short-range" magnocellular dominated rapid detection of motion, and (2) the more complete "long-range" integration of magnocellular motion signals and parvocellular form and position information. In the following discussion the means through which motion-in-depth can be conveyed in this framework will be considered. Earlier, it was suggested that the functional segregation of the visual system is particularly adapted for conveying rapid information regarding potentially threatening objects independent of and prior to a computationally intensive analyses of form. This is particularly true in the case of motion-in-depth and consequently, it can be posited that analysis of such motion-in-depth information can be specified though separate low-level motion processing and higher-level integration of this motion information with featural analyses of three-dimensional structure. David Regan and his colleagues (see Regan, Beverley, & Cynader, 1979; Regan & Beverley, 1979; Regan & Beverley, 1980; Regan, Kaufman, & Lincoln, 1992) have posited such a scheme in which motion-in-depth information is analyzed along the separable pathways of the visual system and subsequently converge upon a unique motion-in-depth stage. This model is based upon psychophysical and physiological data suggesting specific mechanisms for the detection of motion-in-depth.

In psychophysical experiments using adaptation techniques, Regan and his colleagues (see Regan, Beverley, & Cynader, 1979; Regan & Beverley, 1979; Regan & Beverley, 1980; Regan, Kaufman, & Lincoln, 1992) demonstrated the dependence on local directionally selective motion detectors for signaling changing size and motion-in-depth. Regan and Beverley (1980) presented stimuli in which objects exhibited varying degrees of inphase and antiphase oscillation, where pure antiphase oscillations were equivalent to motion-in-depth, and pure inphase oscillations were equivalent to motion along the fronto-parallel plane (see Figure 10). Through these combinations of inphase and antiphase oscillations, motion along various three-dimensional trajectories were simulated. Subjects were adapted to each of these trajectories of motion and subsequently indicated their threshold for the detection of pure inphase and pure antiphase oscillation using the psychophysical method of adjustment. Results demonstrated that threshold elevations for the pure inphase oscillations was functionally determined by both the amplitude of inphase and the amplitude of antiphase oscillation in the adaptation stimulus. However, pure antiphase adaptation produced negligible threshold elevations to inphase test patterns. On the contrary, threshold elevation of the pure antiphase test pattern was independent of the amplitude of the inphase component of the adaptation pattern. From these results, Regan & Beverley (1980) demonstrated that visual responses to antiphase oscillation were independent of the presence of inphase components. Based upon this evidence, it was concluded that the visual system contains changing size channels that receive input from local motion filters that signal x- and y- axis change along stimulus boundaries. These changing size filters consequently can compute an objects rate of expansion or compression which is functionally equivalent to its velocity along the z-axis (cf., Regan, Hamstra, & Kaushal, 1992). In earlier research, it was demonstrated that while pure antiphase oscillations rarely produced a sensation of motion-in-depth, afterimages of antiphase adaptation patterns produced strong percepts of motion-in-depth without any perceived change in object size (see Regan, Beverley & Cynader, 1979). Therefore, it was posited that changing size channels signal motion-in-depth solutions by default unless other cues are available to suggest otherwise. Such an arrangement is clearly adaptive given that motion-in-depth is more probable and more likely to require rapid motor responses such as avoidance (cf., Regan, Beverley & Cynader, 1979).

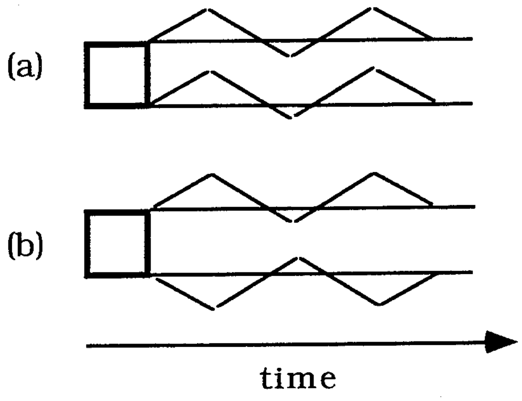

Figure 10. Pure inphase (a) and pure antiphase (b) oscillation of stimulus edges (after Regan and Beverley, 1980).

Regan and Beverley (1979) further demonstrated that adaptation to optical flow patterns indicating motion-in-depth subsequently reduced sensitivity to the changing size of antiphase oscillated test patterns. However, this adaptation to optical flow patterns produced negligible threshold elevations for inphase oscillating test patterns. Furthermore, the degree of threshold elevation among antiphase test patterns was demonstrated to be functionally dependent on the correspondence between the center of the test figure and the focus of the optical flow pattern (see Figure 11). That is, when the antiphase test pattern was presented near the focus of the previously viewed oscillating flow pattern, large threshold shifts were evident. However, when these test patterns were placed at greater eccentricities from the focus of the flow pattern, little or no threshold elevation was evident. This functional relationship between the position of the adaptation flow pattern and the oscillating test object was evidenced in peripheral as well as foveal viewing locations. Based upon these results, Regan and Beverley (1979) concluded that the changing size channels in the visual system are locally sensitive to the region of centrality in an expanding or contracting optical feature. As was already demonstrated in Chapter 1, the importance of the focus of expansion or contraction in conveying the relative direction of travel is a critical invariant for guiding motor actions and egomotion. These psychophysical findings suggest that such an invariant is uniquely represented by locally tuned changing size channels in the visual system.

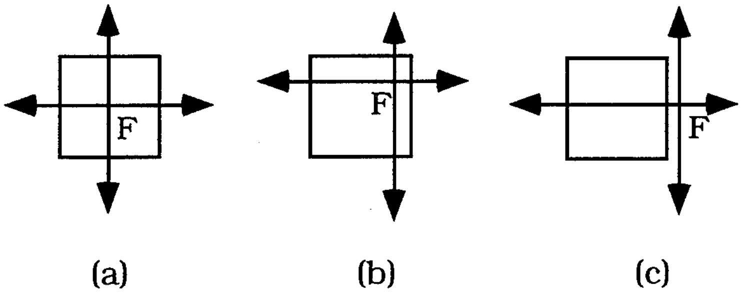

Figure 11. Changing size channels locally tuned to the focus of optical flow patterns. A given channel spatially tuned to the position of the patterns focus of expansion (F) is activated when this point is within a given region as in (a) and (b), but not when it lies outside of this region as shown in (c). (After Regan & Beverley, 1979).

Thus, psychophysical evidence has indicated that information regarding the relative direction of travel in three dimensions is computationally recovered. Lee's (see Lee, 1980) research has further stressed the importance of the relative rate of expansion in providing invariant information regarding the time-to-contact between objects and the viewpoint. Regan, Hamstra, & Kaushal (1992) suggested a scheme in which the interaction of local directionally selective motion filters and changing size filters can signal the rate of an objects expansion. This information can, in turn, be integrated with low level form information regarding the angular extent of the object to compute the relative rate of changing size or tau (τ). A modification of this scheme is presented in Figure 12. According to this model, local directionally selective motion filters signal the direction of motion at object boundaries. The localization of these boundaries can either be estimated by orientation selective spatial filtering (cf., DeValois & DeValois, 1980; DeValois & DeValois, 1988) or can be recovered through the extraction of zero-crossings in a manner such as that suggested by Marr (1982). In any case, a partial analysis of form can suffice in this early rapid analysis since the motion signals are of primary interest. Comparisons of the direction of travel along object borders serve to determine whether fronto-parallel motion or changing size is specified. That is, if motion detection units corresponding with opposite edges of the object signal motion in the same direction then a fronto-parallel motion signal along the x- or y- axis is activated. On the contrary, if motion signals corresponding to opposite edges indicate motion in opposite directions then activation of the local changing size filter occurs. As previously indicated, the changing size filter by default signals motion-in-depth in the absence of contradictory evidence. However, given information regarding object boundaries along orthogonal axes, evidence for motion-in-depth can be augmented by comparison of the rate of change of the stimulus edges along these orthogonal x- and y- axes. In pure motion-in-depth, the rate of change along these orthogonal axes, ΔX and ΔY, are necessarily equal and comparisons that reveal equivalent rates of change provide a strong indication of motion-in-depth. In these cases, the time-to-contact or tau (τ) can be computationally specified by determination of the objects angular extent along either of these orthogonal axes determined from spatial filtering and edge detection (X or Y) and the objects rate of movement along either of these orthogonal axes (ΔX or ΔY). In particular:

$$\tau = \frac{X}{\Delta X} = \frac{Y}{\Delta Y} \qquad (3)$$

Thus, once motion-in-depth has been signaled by changing object size, time-to-contact information specified by the parameter tau (τ) can be directly recovered from lower level computations of local velocity changes and spatial extent. This process clearly requires a symbiosis of motion and form information but clearly can occur prior to exhaustive computations of form information.

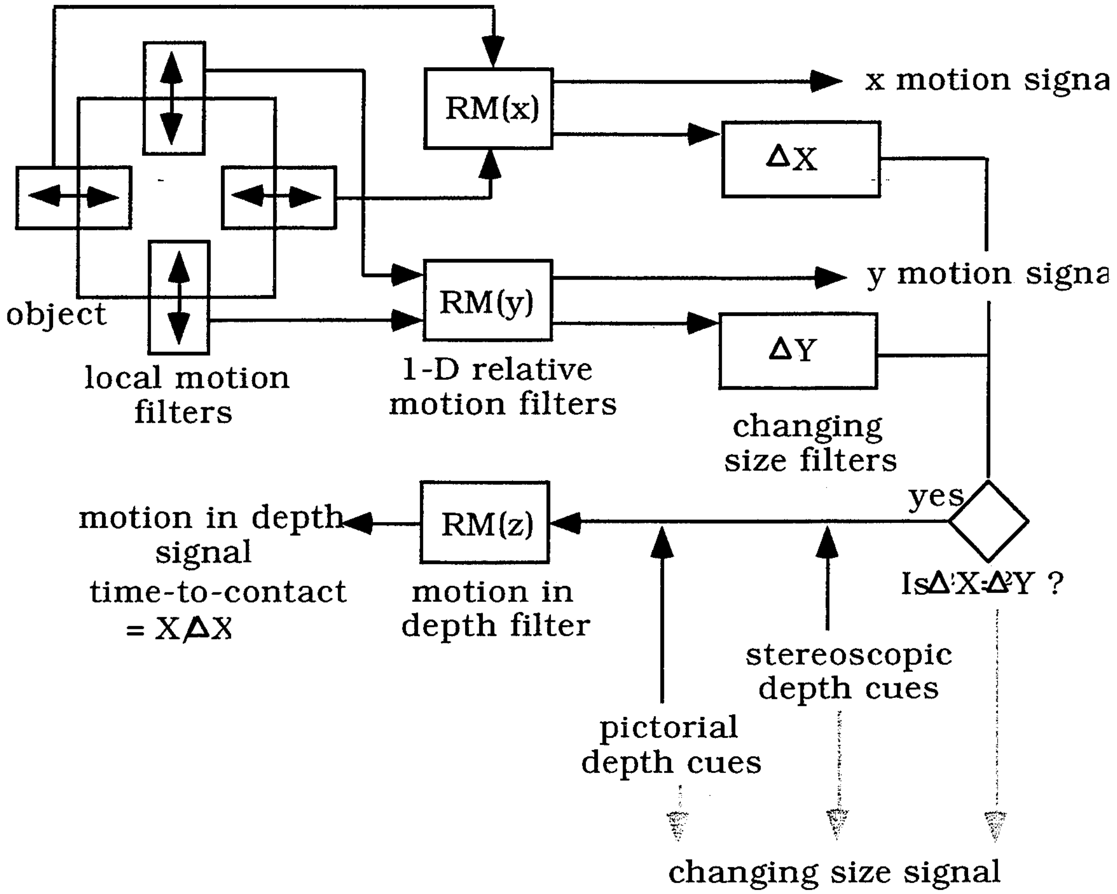

Figure 12. Model of motion in depth processing. Local motion filters signal x- and y-axis motion. Antagonistic motion along opposite edges signal one dimensional changing size. Identical changing size along x- and y- axes indicate either changing size or motion-in-depth. Additional depth cues provide information regarding likelihood of motion-in-depth. If motion-in-depth is perceived, time-to-contact is specified by the relative rate of change in either the x- or y- axis. (Modification based upon Regan, Beverley, & Cynader, 1979 and Regan, Hamstra, & Kaushal, 1992).

In cases not involving pure motion-in-depth, the trajectory of motion in three-dimensional space can be uniquely determined by comparisons of these rates of changes along orthogonal axes. However, the relative time-to-contact with the viewing plane is not easily recovered in such instances. This fact is in accord with psychophysical data that suggests that judgments of times to contact between points along transverse (pure fronto-parallel), oblique (combined fronto-parallel and motion-in-depth) and radial (pure motion-in-depth) trajectories are the products of different computational mechanisms (Schiff & Oldak, 1990). Schiff and Oldak (1990) demonstrated that contact times in transverse motion are the most accurately perceived since they can be computationally recovered from simple analyses of the perceived spatial separation and relative velocity along the fronto-parallel plane. In viewing radial motion in depth, on the other hand, time-to-contact is typically underestimated and more variable. This is not surprising given its reliance upon the determination of higher level expansion information. Furthermore, time-to-contact in such instances is presumably underestimated due to time corrective filters and decision processes that tend to overcompensate for the delay in computing this relative rate of expansion information. Such a conservative corrective process is particularly adaptive since it is typically more detrimental to overestimate the time-to-contact than underestimate it in specifying it as a parameter for motor guidance. Moreover, the parameter is more likely to be underestimated in experimental situations involving discrete motor responses because this time-to-contact information may implicitly contain time based corrections that compensate for the delay associated with parameterizing and executing more complex motor plans. In any case, it is clear that the determination of this time-to-contact parameter in pure motion-in-depth scenarios is dependent on a higher level computation of changing object size. Surprisingly, in cases where travel is along oblique trajectories, accuracy in judging time-to-contact is superior compared to condition in which motion occurs along radial trajectories. Estimation of relative velocities along these oblique trajectories seem computationally complex in that they are only specified through consideration of both the rate of expansion and the rate of displacement along the fronto-parallel plane. However, in these cases, it may be that once a trajectory is determined through analyses of the relative rates of change along opposite edges are determined, a scaling factor can be derived and used to estimate arrival times solely from normalized x- and y- axis displacement values thus decreasing the reliance on higher order changing size information. From these results of arrival time judgments it appears that a model such as the one presented in Figure 12 is a reasonable account for describing how higher order parameters conveying motion-in-depth information can be computationally realized in a biological system.

This psychophysical model of monocular recovery of motion-in-depth information from an analysis of changing size suggests a means through which information regarding time-to-contact can be neurally computed and determined rapidly and early in the visual system (cf., Regan, Beverley, & Cynader, 1979). This monocular sensitivity to changing size and optical flow can be recovered through a partial solution for edge detection such as that provided by the zero-crossings whose computational recovery was described by Marr (1982; cf., Marr & Hildreth, 1980). The motion-in-depth solution from these low-level spatio-temporal characteristics of expanding edges or optical texture features provides a means for the rapid estimation of time-to-contact that can serve to guide responsive motor actions. However, the detection of motion-in-depth and spatial position in depth can be more precisely specified by a more exhaustive consideration of pictorial depth cues as well as static and dynamic binocular disparity cues (cf., Regan, Beverley, & Cynader, 1979; Regan, et al., 1990). The details of these higher level depth and stereo-motion computations will be elaborated momentarily. At this point, discussion will continue along the lines of monocular motion-in-depth specification and the manner whereby form information can be recovered strictly from the relative motion of points will be considered in detail.

Earlier, in discussion of experiments using random dot kinematograms, it was asserted that top-down constraints placed upon the low-level or "short-range" detection of movement can lead to a distinct percept of object form that recovers the boundaries of a displaced region (cf., Sekuler, et al., 1990). Similarly, it has been demonstrated that the systematic motion of points on a two-dimensional projection surface can lead to the recovery of three-dimensional object structure (Ullman, 1979; Hildreth, et al., 1990; see Figure 13). In these demonstrations of the "kinetic depth effect", the recovery of a three-dimensional shape from the systematic displacement of points on a two-dimensional projection is presumed to be the function of the imposition of constraints on the relative motion signals from various points. Ullman (1979) demonstrated that three dimensional shape can always be recovered from the accrual of relative motion information from four points sampled over three frames, and often can be sufficiently specified through the relative motion of three discrete points. Hildreth, et al. (1990), have suggested that even the perception of the relative displacement of only two points can sometimes produce a percept of three-dimensional structure. In the orthographic projection of cylindrical object onto a two-dimensional surface, such as the counter-rotating cylinders shown in Figure 13b, a three-dimensional structure is presumably recovered from the imposition of simplicity and rigidity constraints placed upon the relative motion of points (Ullman, 1979; Hildreth, et al., 1990).

However, these monocular motion cues are insufficient for completely disambiguating depth and consequently reversals will occur in which the cylinders will perceptually change their direction of rotation. Such ambiguities arise from the transparent nature of these stimuli and are rarely observed in the build-up of structure from motion of solid objects. The three-dimensional recovery from motion of such objects can be simulated in computer displays by displacing random pixels in a fashion that mimics the rates of motion of points on a three-dimensional object (see Figure 13a). These displays demonstrate the means through which three-dimensional structure and depth information can be recovered in the environment through the motion parallax produced by movement of structural objects in the environment or movement of the perceiver in relation to structural features in the environment.

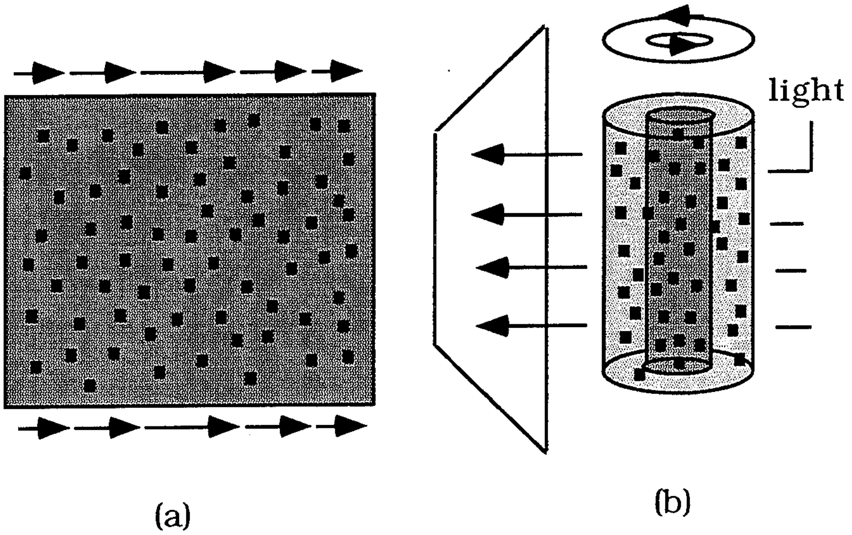

Figure 13. Stimuli in which form can only be recovered from motion: (a) If horizontal motion of the dots is inversely proportional to the vector lengths displayed, a three dimensional cylinder would appear. (b) Orthographic projections of points on two counter-rotating cylinders produces apparent motion on the projection surface from which three-dimensional structure can be recovered. (see Ullman, 1979; Bruce & Green, 1985; Wickens, 1992).

The principal constraint giving rise to this build-up of three-dimensional structure from the patterns of two-dimensional displacements is rigidity (Ullman, 1979; Hildreth, et al., 1990). As Ullman (1979) emphasizes, "Any set of elements undergoing a 2-D transformation as a rigid body moving in space, should be interpreted as such a body in motion" (p. 409). Thus, as with the build-up of two-dimensional structure from motion in random-dot kinematograms through the imposition of rigidity constraints, so too the analysis of three-dimensional structure from relative motion parallax of surface points is constructed through the imposition of higher level rigidity constraints that suggest unique solutions for object structure. As previously suggested, these rigidity constraints must be imposed locally to account for the global elasticity of many real world objects (cf., Johannson, von Hofsten, & Jansson, 1980; Hildreth, et al., 1990). Indeed, the rigidity assumption considers that the transition between discrete points perceptually grouped as components of the same three-dimensional object is smooth and continuous. Consequently, Hildreth, et al. (1990), suggest that the imposition of a rigidity constraint is built-up from comparisons of these discrete points in motion and consequently accrues over time. The manner in which this build up of form information from low-level motion information and high-level object constraints parallels the recovery of structure from motion in the simple case of random-dot kinematograms. Indeed, the two-dimensional solution witnessed in random-dot kinematograms can be seen as a special case of kinetic depth in which the imposition of rigidity constraints lead to the build-up of a form percept of a flat object. It is interesting to note that in the case of random-dot kinematograms, a segregation in depth of the displaced region from the background is often reported. This percept can be the result of the occlusion and shearing of neighboring pixels as the displaced region transverses the random texture array from which it is segregated. The rigidity constraint implies that the object cannot exist on the same depth plane as the texture from which it has been segregated through analysis of structure from motion and therefore comes to be perceived as lying above the texture surface (cf., Gibson, 1979). In the case of three-dimensional recovery of structure from motion, the differential rates of motion of points on a surface presumed to exhibit rigidity allow for estimations of relative depth. As previously noted, these relative depth cues can be ambiguous when viewed in isolation as is the case in Ullman's (1979) demonstration of the orthographic projections of counter-rotating cylinders. In these scenarios, motion parallax provides relative cues to depth of points which allows for recovery of form. However, since absolute depth is not well specified, these demonstrations lead to perceptual reversals of direction due to inherent uncertainty regarding where the points on these transparent surfaces lie in depth. In these cases, the build-up of a three-dimensional spatial representation of object structure and environmental layout can be augmented at higher levels of visual processing that incorporate characteristics of object and texture form into the analysis of depth. Thus, the imposition of rigidity and uniqueness constraints on low-level motion information can produce a initial estimation of three-dimensional spatial layout. This percept of the dynamic three-dimensional layout can subsequently be augmented by a more extensive analysis of available feature information that can provide complementary relative and absolute cues to depth. These monocular or featural cues for depth that cooperatively augment the perception of three-dimensional structure at higher levels of visual information processing are consequently deserving of detailed consideration.

Static Cues for Depth Perception

Monaural pictorial cues to depth necessarily involve at least a rudimentary analysis of object form and textural structure. These cues include: (1) the interposition or occlusion of object regions located at a further distance, (2) the relative size of objects, (3) the relative height of objects in reference to the horizon, (4) linear perspective created by textural features, (5) the foreshortening or decease in size and separation of textural features, (6) the illumination and shading of objects produced by the characteristics of ambient light sources and shadows cast by neighboring objects, and (7) atmospheric perspective or the filtering of light over increasing distance (cf., Goldstein, 1984; Cavanagh, 1987; Landy, et al., 1991; Wickens, 1992). The latter of these cues, atmospheric perspective, while recognized as a contributor to the percept of depth in far-field viewing has not been extensively studied. However, it can presumably be recovered from relatively low level information provided by the spatial filtering of textural and featural components over large scale distances. Similarly, foreshortening can provide cues that can be isolated relatively rapidly through low pass filtering of the luminance characteristics of the dynamic scene. The estimation of distance or depth from foreshortening can be perceptually built-up from analysis of successively finer detail spatial information including the extraction of boundaries or regions of contrast discontinuity in the textural gradient. Similarly, the size and occlusion of objects at a distance can be recovered from a build-up of information regarding their two-dimensional and three-dimensional structure. How these various cues regarding the relative extent and location of features in a two-dimensional projection are joined additively and cooperatively to provide a static monocular percept of three-dimensional layout is of particular interest (cf., Cavanagh, 1987; Landy, et al., 1991).

The manner in which static two-dimensional views lead to the recovery of three-dimensional spatial layout is dependent upon the segregation of form and the imposition of constraints regarding the position of rigid objects within a three-dimensional geometry (cf., Cavanagh, 1987). The manner in which the specific monocular cues specified above are integrated to produce a coherent percept of the three-dimensional structure of the visual scene has been studied in psychophysical experiments where various combinations of cues are viewed in isolation or are placed in conflict with other available cues (see Cavanagh, 1987; Landy, et al., 1991; DeLucia, 1991). For example, Cavanagh (1987) examined the role of occlusion and shading for cueing depth in various images. Findings indicated that in line drawings, occlusion could be recovered in moving, textured, and iso-luminant colored images indicating that the segregation of this static depth information occurred at a high level after these various motion, luminance, and color attributes had been integrated. However, occlusion produced by subjective contours and shadows could only be recovered in instances where luminance contrast information was available, suggesting that these cues may work at lower levels of three-dimensional analysis. These findings suggest that there is an incremental build-up of information regarding geometric position in three-dimensional space provided by the two-dimensional relationships that are conveyed in monocular proximal images. Earlier, the recovery of three-dimensional structure from motion was considered and this too can be seen as contributing to this synthesis of the dynamic spatial layout at various levels in the stream of visual processing (cf., Braunstein, et al., 1986). Consequently, these cues signaled at various levels of processing must be integrated to produce a coherent unitary representation of spatial layout.

Models of depth cue combinatorics posit that the integration of cues resulting from object and observer motion, static monocular or pictorial structural features, and stereoscopic vision involves additive and interactive contributions of these progressively higher level processes (see Landy, et al., 1991). Computational rules specifying the manner in which these cues are combined must be stochastically robust and dynamic in order to address issues of information quality and availability. Various theoretical accounts of how these processes might be computationally realized fall into categories of strong fusion and weak fusion models (Landy, et al., 1991). Strong fusion models contend that depth cues operate primarily as additive factors and are combinatorially weighted to produce a final cohesive representation of depth information. Weak fusion models, on the other hand, suggest that this additive factor process is further augmented by cue interdependencies at various levels of processing. These cue interdependencies include cue promotion and calibration which serve to normalize the information from various stochastic depth cues thereby compensating for the inherent noise and variability introduced in the process of computationally reconstructing the three-dimensional spatial layout. Through cue cooperativity, missing or noisy parameters for a given cue can be estimated by consideration of the depth values obtained from other cues, thereby allowing relative cues to act as absolute cues for distance. Similarly, cooperation between cues can lead to calibration or normalization in which outlying depth indicators are vetoed or devalued in the incremental computation of depth. This further suggests that depth estimation is iteratively updated at progressively higher levels of visual analysis through a system of corrective computations whereby absolute depth is estimated through cue interactions including vetoing, accumulation or additivity, interactive cooperativity, and the disambiguation of cue uncertainty. Through this progressive calibration of depth cue signals, a coherent percept of the three-dimensional spatial layout is dynamically accrued as information from increasingly higher levels of visual processing becomes available.

Landy, et al., (1991) tested the predictions of this theory using a psychophysical method termed *perturbation analysis* in which one depth cue in a pair of available cues is placed in misalignment with the other by some small amount, Δ*cue*. This misalignment (Δ*cue*) is adjusted through psychophysical techniques such as the method of limits, and consequently the perceived change in overall estimated depth (Δ*depth*) is empirically determined through analysis of psychophysical functions of forced-choice responding. From the relationship between the degree of misalignment (Δ*cue*) and the perceived change in distance (Δ*depth*), the relative weighting of the paired depth cues can be determined. Using these techniques, Landy, et al. (1991) demonstrated that kinetic depth from motion and static textural cues interact in a linear fashion in a manner consistent with a weighted average model of depth cue combination. That is, it is apparent that interactions and cooperativity among depth cues is reflective of the relative weighting of these depth cues and this weighting is, in turn, dependent upon the relative quality and consistency of the information provided by each particular cue.

Others have suggested that the combinatorics of monocular depth cues more closely mimics a hierarchical interaction between lower level motion-in-depth cues and higher level static pictorial cues (see, e.g., DeLucia, 1991). The manner in which this hierarchical processing of three-dimensional spatial information distinguishes the interaction of depth cues indicates that the perceptual recovery of spatial position from motion and static cues affords higher combinatorial weighting to cues derived at later stages of computational processing. In particular, DeLucia, (1991) examined the relative percept of time-to-contact when motion based depth information (MDI) obtained from the rate of optical expansion of objects and pictorial depth information (PDI) represented through the relative size of two otherwise identical moving shapes provided conflicting information regarding egocentric distance. Results indicated that judgments of the time-to-contact were underestimated for larger objects and overestimated for smaller objects indicating that time-sampled static cues of relative size were weighted more heavily than the motion based optical expansion cues in estimating arrival time. However, when additional pictorial cues in the form of ground intercept rods, simulated ground shadows, and foreshortened ground reference lines augmented these displays of simulated floating objects, the overestimation of arrival times for small objects was abolished. These findings suggest, that when combinations of higher level static pictorial cues for depth are in accord with lower-level rate of expansion cues a coherent accurate percept of instantaneous position and rate of motion-in-depth can be derived. In this manner accurate spatio-temporal representations of the dynamic scene can be reconstructed and utilized in the parameterization of responsive decisions and motor actions such as estimation of the time-to-contact and execution of appropriate avoidance responses. On the other hand, when higher level pictorial cues conflict with motion-in-depth information, the calibration of dynamic depth information is biased to favor time-sampled comparisons of those higher level pictorial cues even when they are relatively crude. This is somewhat surprising in light of the earlier demonstration that the relative rate of optical expansion is a sufficient parameter for specifying velocity in depth and the remaining time-to-contact between approaching objects and the viewpoint (viz., Lee, 1980; cf., Todd, 1981). The nature of this perceptual weighting bias toward static pictorial cues over motion-in-depth cues might be reflective of the nature of the interaction between low level or "short-range" motion-in-depth signals and higher level "long-range" computations of depth from comparisons of structural features in the visual scene. In particular, the speed and low spatial resolution of motion signals produced by low level processes may be considered to contain a high degree of spatial uncertainty and noise. Consequently, this information is typically augmented through an incremental build-up of higher level, fine detail position information that can serve to normalize the overall dynamic three-dimensional percept in a manner such as that described by Landy, et al., (1991). Under normal viewing conditions, these low level motion signals and higher level static depth cues would typically act cooperatively. However, in anomalous experimental conditions where these motion and position cues are placed in conflict, the low level motion information may be vetoed because it is presumed to be inherently more variable and uncertain.

Thus, from examination of psychophysical studies where various motion and pictorial depth cues are viewed in isolation or placed in conflict, the nature of information accrual regarding three-dimensional layout and the perceptual rules for depth cue combinatorics can be analyzed. The relative contributions of pictorial cues for conveying depth information can be explored further in motor guidance tasks where various combinations of pictorial cues are made available to the operator. For example, in a series of studies, Ellis and his colleagues (see Ellis, et al., 1991; cf., Kim, et al., 1987) had subjects perform three-axis tracking of an object whose dynamic position along these three-dimensions was conveyed through two-dimensional perspective projections. Results indicated that tracking objects in these perspective displays was best at intermediate elevations, near zero azimuthal eccentricities, and when the field of view was relatively narrow. These findings indicate that individuals can recover rather accurate estimates of distance from familiar perspectives, however tracking using solely these monocular cues was demonstrated to be inferior to tracking using stereoscopic displays. Not surprisingly, the earlier described models of depth cue combinations consider stereoscopic cues to be important high level contributors for the computational recovery of a unified percept of the three-dimensional spatial layout. Indeed, the recovery of stereoscopic information can computationally provide absolute indicators of position in depth (cf., Poggio & Poggio, 1984). However, the ability of perceivers to recover this information is complicated by the correspondence that must be achieved between points on the two spatially separated proximal images prior to analysis of the relative disparity between these points. As was the case in resolving the motion *correspondence problem*, theoretical accounts of how this binocular *correspondence problem* is resolved involve the imposition of higher level featural constraints. The specific nature of these theoretical assumptions and the psychophysical and physiological evidence lending credence to these views is deserving of detailed consideration and will serve as the focus of further discussion.

Stereoscopic Vision

The ability to reconstruct the separate proximal images from the two eyes into a single coherent percept of spatial layout is dependent upon the symbiotic analysis of accommodation, vergence, and binocular disparity computations. The latter of these three contributors to the stereoscopic recovery of depth, binocular disparity, will be considered first since it has been the particular focus of computational and theoretical accounts of stereopsis (see, e.g., Marr & Poggio, 1979; Mayhew & Frisby, 1981). Subsequently, the manner in which vergence and accommodation signals may provide key parameters for the determination of disparities between proximal images will be considered. All three of these stereoscopic cues apparently act cooperatively to develop a unique solution to the binocular correspondence problem of assigning correct matches to points on the two proximal images. However, demonstrations of the recovery of perceived depth in the viewing of random dot stereograms through stereoscopic imaging systems indicates that unique correspondence solutions can be derived without vergence and accommodation information, at least over a short range. Furthermore, the recovery of disparity information in these random dot stereograms indicates that correspondences between points on the two proximal images can be recovered without a exhaustive, concomitant analysis of monocular feature boundaries (cf., Regan, et al., 1990).

Random dot stereograms can be constructed by producing two random pixel bitmaps that are identical except for a small azimuthal displacement of a uniform region of the array. When these separate bitmap arrays are viewed independently by the two respective eyes, the degree of disparity, which is functionally dependent upon the degree of displacement of the uniform region, can be perceptually recovered. Consequently, viewing of these displays gives rise to the percept of a fused image in which the binocularly displaced region exists on a different depth plane compared to the rest of the array. This perceptual phenomenon is particularly compelling due to the ambiguity in establishing correspondences between points in the two bitmaps. The correspondence problem introduced in these random dot stereograms is schematically illustrated in Figure 14. In Figure 14, points projected to the right eye (R1-R4) and points projected to the left eye (L1-L4) can be uniquely combined in any of four manners producing 16 potential correspondences of which only four can be correct. In actual viewing conditions, the large number of points whose correspondence must be recovered makes the determination of possible groupings of potential matches highly intractable. Therefore, constraints must be imposed upon the computational processing of these potential matches so that the correct correspondence can be recovered.

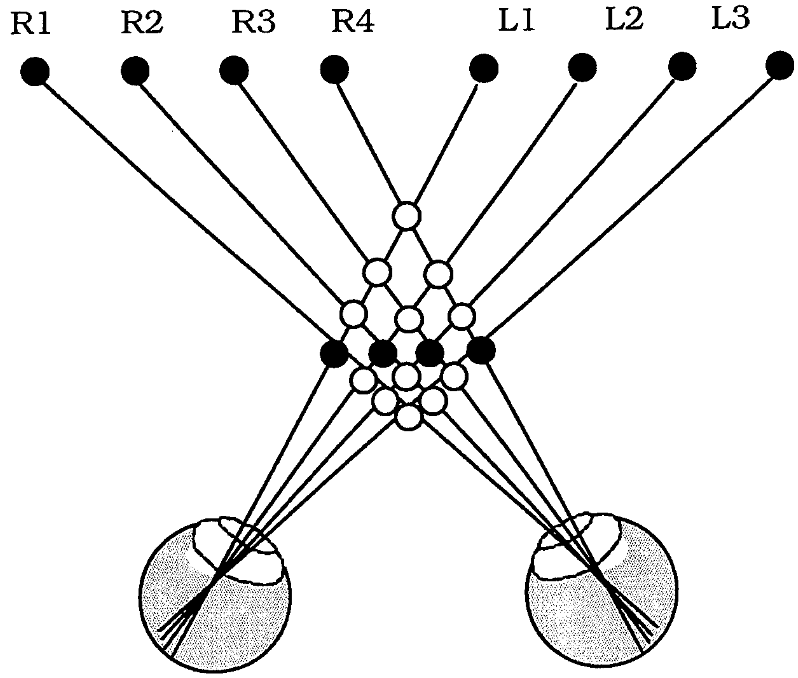

Figure 14. Ambiguity in the correspondence of points on two eccentric proximal images. Permutations of each of the four projections to the two eyes (L1-L4 and R1-R4) can produce 16 potential matches. Of these 16 permutations, only four are correct (filled circles), while the remaining 12 are "false targets". (After Marr & Poggio, 1976; Marr & Poggio, 1979)

Marr and Poggio (1976; 1979) maintain that correspondences are computationally recovered by the imposition of two principal constraints: (1) uniqueness, which indicates that each point may be assigned only one correspondence, and (2) continuity, which maintains that disparity varies smoothly across the surface of a cohesive object. These constraints were discussed earlier in reference to the extraction of edges and the recovery of correspondences between frames in dynamic apparent motion displays. In the case of stereoscopic recovery of disparity, these constraints are imposed in a cooperative matching process that operates locally upon some extraction of rudimentary delineations of form such as *zero-crossings* that serve to demarcate contrast boundaries (cf., Poggio & Poggio, 1984). Consequently, local cooperation of features through imposed constraints of uniqueness and continuity, lead to a rudimentary fusion of corresponding local features into what Marr and Poggio (1979; cf., Marr, 1982) refer to as the *2½-D sketch.* This *2½-D sketch* contains information regarding the surface structure and distance as well as orientations and contours of three-dimensional structure. This information can subsequently be combined with high detail form information derived through feature aggregation processes such as those discussed earlier in reference to form segregation. In this manner, a complete restructuring of the three-dimensional spatial layout can be achieved.

These disparities can be recovered in a biological system through a hierarchical analysis of the locations on the two proximal images where corresponding points lie. This analysis is inherently interdependent with ocular-motor vergence signals and potentially relies on accommodation information since all computed disparities are relative to the point of binocular convergence. This point of convergence is termed the *horopter* or *Vieth-Müller circle* and refers to points in space that produce no binocular retinal disparity (see Patterson & Martin, 1992). Objects aligned a small distance in depth from the horopter fall into a region known as *Panum's fusional area*, so termed because despite the slight disparity between these corresponding points, they are perceptually fused into a single perceived image. Beyond this fusional limit, points are seen as double and the degree of disparity between these points can serve as a cue to depth relative to the point of convergence, the horopter. Objects lying at further distance than the horopter exhibit uncrossed disparities because their corresponding points on the proximal retinal images lie on same side of any referent point cast by an object on the horopter. On the other hand, objects lying nearer in space than the horopter create crossed disparities in which corresponding points lie on the opposite side of the geometrically determined zero-disparity point. Based upon this differentiation of crossed, uncrossed and near-zero disparities it has been suggested that the location of an object in depth relative to the horopter is signaled by the rate of activity among three pools of disparity sensitive mechanisms (see Regan, et al., 1990 for a review;

see Figure 15). Evidence for the existence of these pools comes from samples of stereo-anomalous individuals who are insensitive to either crossed or uncrossed disparities but not both. More conclusive physiological research has indicated that there are cortical cells sensitive to binocular disparity that fall into two distinct classes, (1) narrowly tuned binocular depth cells, and (2) reciprocally activated near and far cells (Ozawa, DeAngelis, & Freeman, 1990; Regan, et al., 1990). The relative activity of these reciprocally activated cells can operate to provide gross distance or depth information regarding crossed and uncrossed disparity points. The more narrowly tuned depth cells, on the other hand, may serve to signal small differences in spatial position around the horopter. In this manner, coarse information regarding stereoscopic distance across the entire range of visible depth and fine information regarding precise estimation of relative distances between objects lying near the horopter can be achieved (cf., Patterson & Martin, 1992). Such a system would clearly represent an economic and efficient strategy for the binocular build-up of information regarding spatial layout derived from successive eye gazes at different vergence points and can also be seen to be particularly adaptive for specifying depth parameters for precision motor responding toward objects lying near the horopter and the provision of precise feedback regarding the outcomes of such actions.

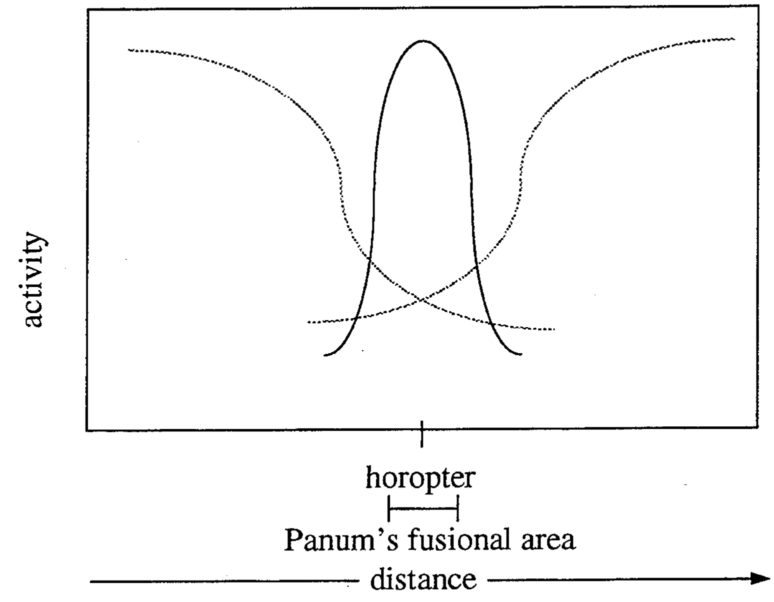

Figure 15. Pools of neuron's sensitive to crossed, uncrossed and near-zero disparity (based on neural characteristics detailed in Regan, et al., 1990).

However, motor actions must often be made in reference to dynamic objects. As described above, stereoscopic information obtained from disparity, vergence, and accommodation information can provide important cues for static position. Additionally, disparity information can provide information regarding motion-in-depth. Earlier, the monocular specification of the rate of expansion was described in detail and it was shown how this low-level process can lead to the recovery of relative rates of motion-in-depth (viz., Lee, 1980; Regan, Beverley, & Cynader, 1979). At higher levels, the relative rate of disparity change between corresponding points can similarly provide sufficient information for the recovery of relative rates of motion-in-depth (see Regan, et al., 1990). It has been demonstrated that if corresponding points on a stereoscopic display are oscillated in an antiphase manner, a percept of motion in depth arises (Regan, Beverley, & Cynader, 1979). Moreover, adaptation to motion-in-depth decreases the sensitivity to detection of these changing disparity patterns suggesting that, in a manner analogous to the monocular recovery of motion-in-depth by the detection of changing size, comparisons of changing disparity can lead to recovery of motion-in-depth (see Figure 12). In particular, Regan, Beverley, & Cynader (1979, cf., Regan, et al., 1990) suggest that four binocular channels for left and right motion at crossed and uncrossed disparities can converge upon two separate motion-in-depth channels, one conveying approach, the other conveying recession of the correspondent object. Specifically, patterns of increasing crossed disparity or decreasing uncrossed disparity signal approach, whereas patterns of decreasing crossed disparity and increasing uncrossed disparity signal movement of the object away from the viewer. Thus, this higher level binocular determination of motion in depth can serve to cooperate with lower level monocular determination of expansion rate to produce a coherent dynamic spatial percept of an object in motion.

This interaction of monocular and binocular cues to motion-in-depth serves nicely to recapitulate the theme of this chapter which posits that the perceptual recovery of the dynamic spatial layout of a three-dimensional scene involves cooperative computation of information at progressively higher levels of processing. In this manner, rapid motion information and rudimentary form information can be provided for the parameterization of rapid response plans, and this information can be progressively augmented through more elaborate computations of structure and three-dimensional position thereby permitting the reconstitution of response plan parameters to allow for precision in motor actions. Thus, the coherent perception of dynamic spatial layout involves a complex, symbiotic coordination of the various pathways and mechanisms in the visual system specifically adapted to carrying out highly specialized tasks such as the detection of motion and the analysis of fine detail. However, vision is not the only modality capable of conveying spatial information. Hence, the focus of this inquiry will now turn to a consideration of the perception of dynamic spatial layout through information processed by the auditory modality. In this analysis of auditory spatialization, it will be demonstrated that, unlike vision, the recovery of auditory motion and position information is highly interdependent, a consequence of the fact that acoustic signals are inherently non-spatial in character thereby rendering the recovery of auditory spatial information dependent upon a concomitant analysis of acoustical structure.

AUDITORY SPATIALIZATION

Auditory spatialization refers to the ability to judge the direction and distance of an environmental sound source or the percept of direction and distance achieved through headphone simulations of three dimensional auditory spatial cues. (Moore, 1988; Wightman & Kistler, 1989b). Auditory spatialization can be further extended to include the detection of motion of auditory sources including the perception of velocity and acceleration of sound sources in azimuth, elevation, and depth. The perception of auditory spatial location and spatial dynamics is a complex function of characteristics of the sound source, the acoustical environment, and the physiological and anatomical characteristics of the perceiver.

The Perception of Acoustical Structure

The examination of auditory spatialization must begin with a consideration of how sound sources come to be perceived as distinct physical entities. Unlike vision, the auditory signal transduced by the perceiver is inherently non-spatial in character. That is, the sound signal reaching the ears of the listener simply reflects amplitude and frequency changes over time. It is only through an analysis of the sound spectrum that spatial information can be recovered. However, extraction of this dynamic spatial information is dependent upon segregation of the auditory stream so that sounds emanating from distinct sources can be isolated from the ubiquitous acoustical ambience present in the environment. In fact, little is known about the potential means through which sound sources come to be distinguished. However, several organizational strategies appear to be employed in parsing the stream of auditory information. These organizational strategies revolve around the extraction of temporal patterns and rhythmic regularities in the auditory stream. These elements of the temporal pattern consequently can be parsed by consideration of constraints and relationships defined by features of the acoustic signals. These constraints and relationships were recognized by the Gestalt psychologists and include structural similarity, good continuation, common fate, belongingness, and closure (Moore, 1988). These structural relationships can emerge from temporal analyses of correlated changes in the sound spectrum. For example, sounds of similar frequency and amplitude will tend to be perceived together. Moreover, since changes in frequency and amplitude emanating from a single sound source are typically smooth and relatively continuous, large differences in frequency and intensity can serve to isolate acoustical sources. Likewise, component frequencies emanating from a single source will demonstrate common temporal changes and hence are perceptually joined by common fate. Providing further constraint upon the problem of segregation is the simple matter of belongingness, or the fact that a frequency component can only arise from a single sound source. However, this provides potential difficulties in that sources of narrow bandwidth can potentially be masked by the surrounding auditory ambience. Despite potential masking, sounds that are briefly obscured can be continuously perceived through a filling-in of the masked signal based upon the assumption of closure and good continuation (Moore, 1988; Deutsch, 1975; Deutsch, 1986).

The other major source of information regarding form is reflected through spatial differences. Therefore, auditory analysis of form identification and spatial information must be performed concurrently since they are inherently interdependent. This is evidenced in the release from masking of sounds exhibiting overlapping spectral characteristics with increasing separation between acoustic sources (see Moore, 1988; Doll, Hanna, & Russotti, 1992; Good & Gilkey, 1992). This release from masking is also evidenced in experiments using binaural presentations of signals and noise. In particular, release from masking can occur when a binaural tone masked at threshold by noise is phase shifted by 90 degrees in one ear. More compellingly, release from masking can also occur when noise is added in the opposite ear of a monaural tone masked at threshold. As will be discussed momentarily, these phenomena arise from the introduction of interaural differences that provide resolvable cues for localization. The resulting ability of humans to utilize these localization cues to distinguish auditory objects from the surrounding acoustic ambience has often been referred to as the *cocktail party phenomenon* because of its functional role in allowing listeners to follow a single conversation in a room filled with ambient speech signals. However, the requirement for concurrent resolution of information regarding structure and location of auditory signals can lead to perceptual ambiguities. For example, Deutsch and her colleagues (see Deutsch, 1975; Deutsch, 1986 for reviews) have demonstrated that localization of concurrent signals can be made ambiguous by spatially dissociating organized temporal patterns. For example, simultaneous ascending and descending musical scales presented dichotically can appear to remain localized in one ear even when they are spatially reversed in the midst of presentation. Such perceptual illusions stress the critical fact previously stated that acoustical form and localization are perceptually determined concurrently and interactively.

Thus, form and location cues are simultaneously analyzed and synthesized from the neurally transduced sound spectrum representation of the ambient auditory stream. This ability to use localization cues concurrently with cues obtained from temporal patterning of the sound spectrum allows perceivers to reconstruct a complex spatio-temporal representation of acoustic objects and events. Such a representation can convey information about the spatial and temporal dynamics of complex task situations important for cognitive decisions and motor response planning (cf., Folds, 1990; Gaver, Smith & O'Shea, 1991). Accounts of how neural processes give rise to such a dynamic spatial representation focus on the acoustical cues available to the listener. Furthermore, the manner in which these localization cues are temporally encoded has been a major focus of auditory perceptual research. Historically, localization was thought to be principally conveyed through time-sampled binaural comparisons such as those suggested by the masking phenomena discussed above. More recently, the importance of monaural cues created by distortions of the sound signal as it travels through the environment and is perturbed by anatomical structures of the listener such as the pinna has been stressed. The specific nature of these various binaural and monaural spatial cues will be considered in detail in the following discussion.

Duplex Theory of Auditory Localization

The ability to localize auditory stimuli in azimuth and elevation is significantly impacted by spectral properties of the sound source. Thus, large differences in auditory spatial acuity are found in comparing empirical results from studies using pure tones, beats and clicks, broadband noise, and spectrally filtered noise (Middlebrooks & Green, 1991). However, theoretical accounts of the perception of location among these diverse acoustic stimuli posit three critical cues for providing auditory localization information: (1) interaural intensity difference cues, (2) interaural phase or time difference cues, and (3) anatomical distortion cues created by the pinna, head, and upper torso of the perceiver. The importance of the first two of these cues was recognized by Lord Rayleigh in the late 19th century and are described in his dual-process or "duplex" theory of horizontal sound source localization (see Rayleigh, 1907). These two processes originally described by Lord Rayleigh involve the neural computation of (1) interaural intensity differences and (2) interaural time or phase differences.

Interaural intensity differences refer to the absorption of sound by the head leading to sound pressure differences of stimuli reaching the two ears. The result is a sound shadow of attenuated sound pressure levels reaching the ear further from the sound source (Moore, 1988; Handel, 1989). This sound shadow is a complex function of the spectral characteristics and position of the sound source relative to the listener. Interaural intensity differences are negligible below 1,500 Hz but can be as great as 20 dB at frequency levels above 5,000 Hz (Moore, 1988). At high frequencies, the interaural intensity differences follow an inverted-U function in relation to the sound source's azimuthal angle of incidence with the listener's head.[3] This inverted-U pattern of interaural intensity differences at high frequencies exhibits a peak when the sound source is located near 90 degrees in azimuth. Actually, the inverted-U pattern is a generalized oversimplification of the effects of sound source azimuth on interaural intensity differences. In fact, the pattern of interaural intensity differences demonstrates a peak at points slightly eccentric from 90 degrees, and a small local minima occurs at 90 degrees. This dip in the pattern of interaural intensity differences at 90 degrees occurs because of the convergence of the diffracted sound waves by the head which partially rejoin on the opposite side of the listener's head to form a localized region of summated sound pressure levels termed the *central lobe*. Since the head is roughly spherical, this *central lobe* is localized at the position of the far ear when the sound source is at approximately 90 degrees in azimuth and therefore serves to slightly offset the interaural intensity differences in such cases (see Handel, 1989). In any case, sound pressure level differences between the near and far ear are substantial at high frequencies. However, within lower frequency ranges, localization is presumed to be mediated by interaural phase or time differences.

[3] Azimuthal incidence angle is measured in reference to the medial sagittal plane (passing through the midline of the nose). Sound sources occurring along this plane in front of the listener are assumed to have an incidence angle of zero degrees in azimuth, and sound sources occurring directly behind the viewer along the medial sagittal plane have an incidence angle of 180 degrees. Assuming symmetry, incidence angles are described as ranging between zero degrees and 180 degrees.

Interaural phase or arrival time differences reflect the fact that it takes a fixed amount of time for a sound wave to travel a certain distance such as the distance separating the two ears. This implies that at any given point in time the sound wave from a single source impinging upon the two ears is in a different phase (assuming, of course, that the wave impinging upon the ears is not phase shifted by 360 degrees). It further implies that a fixed period of time intervenes between the time when the sound impinging on the nearer ear is in a certain phase alignment and the time that the sound reaching the further ear comes into that same alignment (cf., Moore, 1988; Handel, 1989). The interaural phase difference can be assessed by determining interaural path differences, the difference in distance traveled by the sound source as a function of incidence angle, which can be represented by the following equation:

$$d = r\theta + r\sin\theta \qquad (4)$$

where *d* is the interaural path difference, *r* is the radius of the listener's head, and is equal to π minus the incidence angle of the sound source in radians (refer to Figure 16; Moore, 1988). Thus, the interaural time difference, *T*, can be derived by dividing the obtained interaural path distance, *d*, by the speed of sound, *c*, (cf., Middlebrooks & Green, 1991):

$$T = \frac{d}{c} = \frac{r}{c}(\theta + \sin\theta) \qquad (5)$$

In this manner, interaural time differences can be approximated as a function of incidence angle assuming a perfectly spherical head and 180 degrees separation between the two ears. Furthermore, interaural phase differences can be determined as a function of incidence angle and frequency. These derived interaural time differences are quite small, on the order of less than 1 ms, but nonetheless are perceptible for auditory frequencies below approximately 1,400 to 1,600

Hz (Scharf & Houstma, 1986; Moore, 1988; Handel, 1989; Middlebrooks & Green, 1991). As with interaural intensity differences, interaural time or phase differences within this range can be represented by a symmetric inverted-U function peaking at 90 degrees in azimuth (Scharf & Houstma, 1986; Handel, 1989). At zero degrees and 180 degrees the path difference between the ears as defined by Equation 4 is zero and consequently no interaural time or phase differences are present along the medial sagittal plane. For sounds directly on the side of the listener, at 90 degrees in azimuth, interaural time differences range between 0.6 and 0.8 msec (Scharf & Houstma, 1986; Handel, 1989).

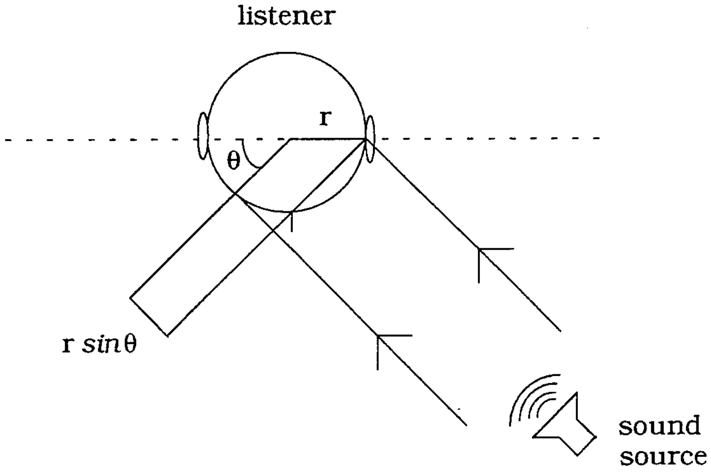

Figure 16. Schematic of the interaural path difference computed by Equation 1 assuming a perfectly spherical head and 180 degrees separation between the two ears (Moore, 1988).

As previously mentioned, empirical evidence has demonstrated that interaural phase differences are perceptible in the range below 1,400 to 1,600 Hz. Above this range, phase differences of 360 degrees can occur and lead to ambiguity regarding which ear is receiving leading information (Middlebrooks & Green, 1991). However, above approximately 750 Hz, phase differences of 180 degrees between the two ears can occur given that the maximal path difference between the two ears is approximately 23 cm which is equal to the half-wavelength of a

750 Hz tone (Moore, 1988). When the sound impinging upon the two ears is phase shifted by 180 degrees, location is similarly ambiguous for continuous tones given the uncertainty regarding which ear is receiving leading information. These ambiguities can be resolved through head movements that change the azimuthal incidence angle and consequently the path difference between the ears. Thus, sampling the sound source at different incidence angles can serve to disambiguate phase relationships between the ears. Furthermore, the presence of transients, such as clicks, onsets, and offsets, can provide interaural time cues that are not confounded by lead ambiguities in phase relationships (Moore, 1988).

Thus, below 750 Hz, interaural phase differences can accurately provide localization information regarding continuous tones to stationary listeners. Between 750 Hz and approximately 1,500 Hz, sampling of interaural phase differences at different incidence angles or detecting interaural time differences in the onsets and offsets of transients can provide accurate cues for localization. Above 1,500 Hz, interaural time and phase differences are generally imperceptible despite evidence that auditory phase coding has been demonstrated in other primates at frequencies as high as 5,000 Hz and in barn owls at frequencies as high as 7,000 Hz (Middlebrooks & Green, 1991). Therefore, it can be concluded that this upper limit on interaural phase and time differences are not determined by limitations in neural temporal patterning. Nonetheless, despite an inability to perceive phase lags at frequencies above 1,500 Hz among human listeners, higher frequency sounds that are amplitude modulated can yield perceptible interaural delays (Moore, 1988; Middlebrooks & Green, 1991). These amplitude modulated delays can provide spatial information, but sensitivity to these interaural time differences is diminished when carrier frequencies rise above 4,000 Hz or when modulation frequencies rise above 500 Hz (Middelbrooks & Green, 1991).

From these general observations, the "duplex" theory of horizontal sound source localization has posited that interaural intensity differences operate at frequencies above 5,000 Hz, and interaural phase differences operate at frequencies below approximately 1,500 Hz. Since the auditory system is presumed to analyze complex sounds by decomposing them into their fundamental sine wave components, each of which is distinctively perceptible, both interaural intensity and interaural time or phase differences can be utilized for localizing broadband acoustic signals (cf., Moore, 1988). The presence of transients and amplitude modulations can provide localization cues between 1,500 Hz and 5,000 Hz.

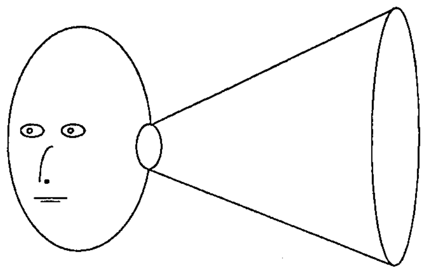

Figure 17. The auditory cone of confusion describing points in space that are mathematically determined to give rise to the same interaural path, intensity, time, and phase differences based upon the assumption of a perfectly spherical head (cf. Moore, 1988; Handel, 1989).

However, the duplex" theory of neural computation of intensity and phase or time differences is an incomplete account of auditory localization for several reasons. First, it fails to adequately describe horizontal sound source localization abilities across the full range of auditory frequencies perceptible by human listeners. Second, the azimuthal resolution obtained by providing just intensity and phase difference cues through headphones is often larger than that obtained in free-field listening conditions. These qualitative differences between headphone presentations of interaural cues and free-field listening have served to distinguish between auditory *lateralization* and auditory *localization* (see Moore, 1988; Wightman & Kistler, 1989a and 1989b). This distinction between lateralization and localization refers to the fact that sounds presented stereophonically through headphones tend to be perceived within the head and lateralized between the two ears rather than being perceived as localized in the external environment at some distance from the listener. Third, interaural differences fail to provide adequate cues for localizing sound source positions in the vertical axis. In fact, along the medial sagittal plane, no interaural cues are available for distinguishing vertical position. In general, the poor resolution of localization based upon interaural cues has been described as giving rise to a *cone of confusion* (see Figure 17). Any sound emanating from a point of the surface of such a cone will give rise to the same interaural intensity and phase differences (see Handel, 1989). Since mathematically constructed cones of confusion are much larger than empirically observed spatial acuity measures of sound source azimuth and elevation, it is apparent that other cues are available and are utilized in localizing sounds. In fact, the distortion of sound by anatomical structures of the listener, such as the pinna and upper torso, provides important cues that minimize this cone of confusion thereby improving spatial acuity in the auditory modality.

The Role of Anatomical Transfer Functions

As sounds are transmitted through the air, they are diffracted and absorbed by many environmental objects including anatomical features of the perceiver. The critical role of sound absorption by the head in creating perceptible interaural differences has already been considered. However, the distortion of sound produced by other anatomical structures of the listener provides critical cues for localization, particularly along the vertical axis and for sounds lying along the medial sagittal plane. Of particular importance is the structure of the pinna or outer ear which has adaptively evolved to provide an important transformation of the sound spectra of acoustic stimuli impinging upon the ear (see Batteau, 1967). The convolutions or folds of the pinna allow for multiple paths through which the sound wave can travel. These paths serve to transform or filter the sound into time-delayed replications or reverberations. Thus, the pinna has been described as acting as a "comb-filter" that produces a transformed, time-delayed frequency spectrum with peaks and valleys (see Batteau, 1967; Moore, 1988; Middlebrooks & Green, 1991). This modification of the waveform impinging upon the inner ear has a significant effect on the ability to localize sounds both in azimuth and elevation. In terms of azimuth, the cues provided by the pinna are particularly important in determining the location of sounds lying on the medial sagittal plane and consequently eliminating front-to-back confusions regarding sound source origin. However, research investigating pinna cues using headphone presentations of recorded sounds filtered by artificial pinna molds and unfiltered sounds has demonstrated that the inclusion of pinna cues significantly improves localization ability across the entire range of azimuthal positions (Batteau, 1967; Moore, 1988; Middlebrooks & Green, 1991). The improved spatial acuity effects of pinna cues on vertical sound source localization ability is even more pronounced. Moreover, presentations including pinna cues are qualitatively described as giving rise to the sensation of *localization* outside of the head rather than *lateralization* within the head between the two ears (Moore, 1988; Wightman & Kistler, 1989). Finally, the cues provided by the time-delayed filtering of the pinna allow for monaural sound source localization which can be contrasted with the binaural neuronal comparison processes that account for interaural intensity and time or phase differences (see Butler, 1987).

The perceptibility of the monaural time-delay cues provided by the pinna is limited by the ability to separately encode the principle signal and its time-delayed replications. The lower limit on this encoding in human listeners has been shown to be about 6,000 Hz. Therefore, pinna filtering cues of the acoustic wave appears to be a significant cue for sound source localization at relatively high frequencies (see Batteau, 1967; Moore, 1988). However, some research has demonstrated that these pinna filtering cues can be utilized at frequencies as low as 3,000 Hz (see Moore, 1988). Nonetheless, monaural localization of low frequency sound is generally quite poor, especially in reference to vertical position. This fact also suggests that other anatomical features, namely components of the head and upper torso, may play a role in the transformation of sound source spectral profiles, especially at lower frequencies.

Besides the pinna, the reflectance of sound waves by the listener's upper torso appears to provide significant cues to sound source location (Wenzel, 1991). The reflectance of sound from these surfaces is not fully understood, but their importance can be ascertained indirectly through empirical determination of head-related transfer functions (HRTFs) of the sound source. These HRTFs can be obtained by inserting recording microphones at the opening of the inner ear in either human listeners or head mannequins using pinna molds of human ears (Butler, 1987; Moore, 1988; Wightman & Kistler, 1989a; Wenzel, Wightman, & Kistler, 1991). The transfer function can then be determined as the difference between the sound spectrum at the acoustic source and the sound spectrum at the inner ear microphone assuming that a controlled acoustic environment such as an anechoic chamber is used to prevent confounding transformations of the acoustic signal by environmental objects. The observed anatomical filtering is largely resultant from the shape of the head and pinna but can only be completely described by consideration of the entire upper torso of the listener. Indeed, an accurate modeling of the attenuation of sound pressure levels by the head is most critical for describing the sound shadow that gives rise to the perceptible interaural intensity level differences previously discussed. Similarly, modeling of the interaural paths about the listener's head is critical for accurate descriptions of interaural differences. That is, Equations 4 and 5 assume a perfectly spherical head and therefore introduce a degree of error in computations of interaural phase differences. Actual interaural differences can be more accurately measured through empirical testing with actual human listeners or accurate models of human heads with attached model pinnas. Transfer functions that fail to capture the role of these anatomical structures such as the pinna, head, and upper torso demonstrate poorer localization judgments when presented to human listeners than transfer functions that accurately account for the distortions of the sound spectral profiles by these features (cf., Butler, 1987; Wightman & Kistler, 1989a).

With the advent of high-speed digital signal processing (DSP) computer hardware, these HRTFs can be used to digitally filter sound signals and present them to listeners in real time while compensating for head movements measured by head tracking devices (Wightman & Kistler, 1989a and b; Loomis, et al., 1990; Begault & Wenzel, 1991; Wenzel, 1991; Wenzel, Wightman, & Kistler, 1991). In this manner headphone simulations of environmental auditory spatial cues have be achieved with some degree of success. Localization abilities using these virtual acoustic displays have be shown to be promising yet not on the same level as spatial acuity for environmental sound sources (Wightman & Kistler, 1989b). It has further been demonstrated that using non-individualized HRTFs can lead to fairly good localization, however, results in these conditions are somewhat inferior to results obtained with HRTFs based on the specific listener's anatomy (Wenzel, 1991; Wenzel, Wightman, & Kistler, 1991). Besides the general decrease in acuity with these displays, front-to-back reversals are quite common thus suggesting that these displays fail mainly due to the fact that the HRTFs don't adequately capture the complex transformations carried out by the pinna and the upper torso.

Static Auditory Acuity

In the preceding discussion, the means through which the distortions of the sound signal created by the anatomical characteristics of the perceivers head, pinna, and upper torso come to serve as monaural and binaural cues to sound source location has been described in detail. Through psychophysical experiments in which specific transfer characteristics are approximated through headphone presentations, the relative influence of each of these transformations can be determined (cf., Middlebrooks & Green, 1991). That is, by introducing signal intensity, phase, or time differences in dichotic headphone presentations and measuring position estimation error and variance, the contribution of each of these cues for sound localization can be ascertained through comparisons with free-field measures of auditory spatial acuity. However, these techniques typically underestimate the role of these cues, especially when an assumption of a perfectly spherical head is made. The relative impact of pinna cues can similarly be assessed through remote headphone localization estimations of sounds passed through individualized and non-individualized pinna molds (Batteau, 1967; Wightman & Kistler, 1989; Wenzel, Wightman, & Kistler, 1991). However, all of these cues individually fail to adequately describe static spatial acuity abilities in the auditory modality. Thus, psychophysical determinations of spatial acuity in free-field listening have been employed to capture the additive and interactive influences of these anatomical transformations of the sound signal for auditory localization.

The psychophysical methods employed in evaluating acoustic spatial ability can be grouped into two general categories: (1) measurement of error and variance in pointing responses, and (2) threshold detection of the minimum perceptible angle of sound source displacement or the minimum audible angle (MAA). The pointing method is frequently employed because it is quite straightforward and amenable to rapid data collection. Pointing responses can be measured through simple hand positioning, or through more complex measures of orienting such as eye or head tracking (see, e.g., Sorkin, et al., 1989). Measures of absolute and relative errors in azimuth and elevation and the correlation between actual and reported sound source position can be derived in this manner. Data using these methods have been collected under both free-field listening conditions and headphone simulations (e.g., Sorkin, et al., 1989, Wightman & Kistler, 1989b; Makous & Middlebrooks, 1990). General results are a complex function of azimuth, elevation, acoustic characteristics of the sound source, and mode of presentation (free-field versus individualized virtual displays versus non-individualized virtual displays). Correlations between actual and reported azimuthal position tend to be quite high, but decrease significantly as the source becomes more peripheral. Elevation judgments are typically much poorer than azimuth judgments and are particularly poor along the medial sagittal plane (Butler, 1987; Wightman & Kistler, 1989b; Makous & Middlebrooks, 1990; Wenzel, Wightman & Kistler, 1991). While these methods readily allow for comparison of auditory spatial abilities between free-field listening and headphone simulations, they are inadequate for determining the acuity of auditory spatial localization. To determine localization acuity in the auditory modality, methods for determining MAA are typically employed. These methods assess the smallest separation in auditory sound sources that can be reliably detected using standard threshold detection tasks (see Moore, 1988). These studies have shown that directly in front of the listener, MAA can be as small as one degree but increases to as much as 40 degrees at 90 degrees of azimuth for certain sounds (Strybel, Manligas, & Perrott, 1992). While these results indicate that auditory spatial acuity can be quite good, at least directly in front of the listener, this acuity is still much poorer than acuity in the visual modality which can be smaller than one arc minute of visual angle for high contrast stimuli (Wilson, 1986). The impact of this difference in acuity on tasks in which visual and auditory information must be compared and subsequently pooled in order to specify guidance parameters for motor actions remains to be determined.

Auditory Motion Detection

Of more particular relevance to motor guidance, is the dynamic spatial acuity of auditory sound sources. The ability to detect motion in the auditory modality has been measured by determining the minimum angle of displacement required for reliable detection of the direction in which the sound is traveling: the minimum audible movement angle (MAMA; see, e.g., Perrott & Marlborough, 1989; Strybel, Manligas, & Perrott, 1992). Detection of motion is a complex function of azimuth, elevation, sound source acoustic characteristics, and sound source velocity. Unfortunately, the latter of these determinants, sound source velocity, has not be systematically varied in the course of a single experiment (cf., Strybel, Manligas, & Perrott, 1992). Nonetheless, MAMA values of approximately one to three degrees at zero degrees in azimuth, and seven to ten degrees at extreme peripheral locations (near 90 degrees in azimuth) have typically been found (Perrott & Marlborough, 1989; Strybel, Manligas, & Perrott, 1992). Similarly, obtained MAMA values as a function of elevation typically demonstrate increases with decreasing centrality of the sound source. At 90 degrees of elevation, MAMA increases to about seven degrees (Strybel, Manligas, & Perrott, 1992). The similarity between MAA values and MAMA values and the fact that both are influenced by the same external factors leads one to believe that these acuity measures are the result of the similar neural processes. That is, dynamic acuity (MAMA) might well be the result of time sampled location estimates that are described by measures of static spatial acuity (MAA).

With increasing time available to sample the spatial dynamics of a sound source, highly accurate estimates of velocity and instantaneous position can be attained. The ability of the auditory modality to accurately perceive the velocity of sound sources was demonstrated by Waugh, Strybel, and Perrott (1979). In a set of experiments, Waugh, Strybel and Perrott (1979) demonstrated that judgments of the absolute velocity of a moving sound source were as accurate as judgments made visually across a wide range of object velocities despite the significantly poorer static spatial resolution of the auditory modality as compared to the visual modality. In subsequent research, Perrott, Buck, Waugh and Strybel (1979) demonstrated that velocity estimates for moving sound sources were significantly affected by the listening duration. For extremely short listening periods (i.e., 100 msec or less), velocity was typically underestimated, and frequently no motion was perceived. However, for listening durations of 300 msec and greater, highly accurate estimates of velocity were obtained. Results from these studies consistently produced nearly a one-to-one correspondence between the actual and perceived velocity of the sound source when listening times exceeded approximately 300 msec. These findings can be accounted for by a process of stochastic estimation based upon the incremental accrual of time-sampled position estimates of the sound source. This theory of stochastic estimation of time sampled sound source position estimates was first presented in my dissertation (Elias, 1994) and is recapitulated below in the context of the present discussion.

After the onset or initial detection of a dynamic sound source, uncertainty regarding the position and velocity of that source rapidly diminishes. With increased time to sample the instantaneous position of the sound source, spatial ambiguity can be further resolved and uncertainty can be further reduced. This occurs for two principle reasons. First, sampling of the sound source over multiple positions provides more data points for perceptually computing velocity estimations. Second, multiple samplings reduce the biasing effects of noise and equivocation through stochastic summation of instantaneous position and velocity estimates. That is, by increasing the time available to sample the sound, better estimations of position and velocity through the summation of time sampled location estimates is possible (see Figure 18).

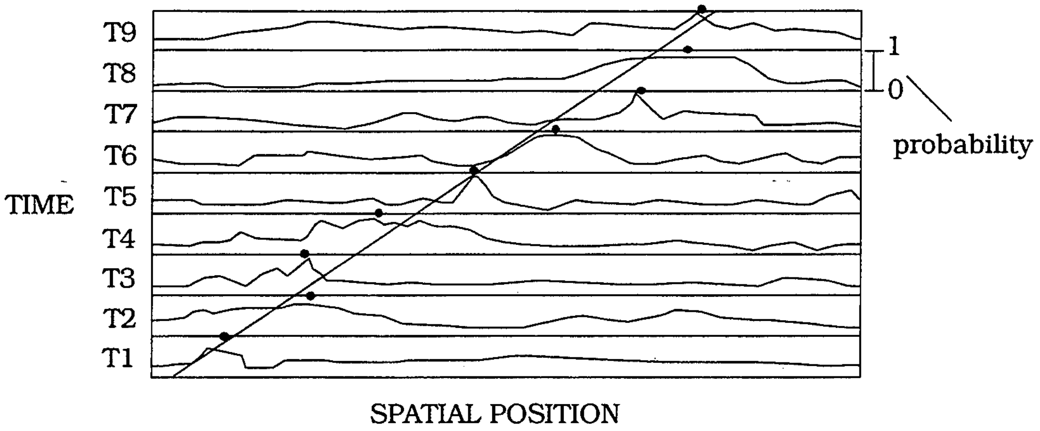

Figure 18. The process of stochastic estimation of sound source position and velocity. The graph depicts hypothetical probability estimates of spatial position across time sampled observations. From the probability distributions, maximum likelihood estimates of spatial position can be determined and are indicated by solid circles. Using error minimization techniques, a best fit estimate of velocity can be computed and is indicated by the slope of the solid line.

In Figure 18, hypothetical time sampled probability estimates of spatial position are plotted. Each estimate is representative of the instantaneous spatial uncertainty that a listener would have based upon a single sampling of the localized sound source. On certain sampling occasions, noise and equivocation of the signal may produce high uncertainty. Nonetheless, the listener could compute an estimate of the sound source location from the distribution of instantaneous position likelihoods. Initially, this computation would be largely independent of other samplings. However, later estimates may be weighted or biased by previous estimates by incorporating assumptions regarding the smooth continuous motion of the sound source through space. In this manner, best estimates of sound source position could be derived for each temporal sampling of position likelihood. These best sampling estimates are represented by the solid circles in Figure 18. Finally, using estimation techniques to derive a best fit for the distributions of estimated spatial position over time, estimations of dynamic spatial properties of the sound source can be determined. Again, assumptions regarding the characteristics of sound source travel would greatly influence these computational procedures. For example, if it were suggested that the sound source was accelerating or decelerating, computational procedures could be biased toward fitting a power function to the change in estimated sound source position. On the other hand, in the experiments discussed above (Waugh, Strybel, & Perrott, 1979; Perrott, et. al., 1979), the sound source was known to travel at constant velocity. In such instances, computational procedures for estimating velocity might initially proceed by deriving a linear fit to the distribution of estimated sound source positions. In this manner, the estimated velocity can be determined by the slope of the best fitting linear estimation of sound source position over time (see Figure 11). Consequently, accurate estimations of sound source velocity can be recovered even when spatial acuity is rather poor or aspects of the task environment render precise localization of the sound source quite difficult. However, increasing the number of time samples would produce decreasing marginal gains in the reduction of error between the actual and estimated paths of sound source travel. From the findings obtained by Perrott, et al. (1979), it appears that beyond 300 msec of listening, the marginal gain in the accuracy of velocity estimations is insignificant. Furthermore, below approximately 100 msec insufficient time samples have been acquired for accurate velocity estimations. Therefore, this process of improving velocity estimations through time sampled corrections appears to operate primarily in the range between 100 and 300 msec. While this is a sufficient account of angular velocity estimations acquired through the auditory modality, it fails to address the perception of static and dynamic depth characteristics of sound sources. In order to more fully comprehend the spatial perception of auditory stimuli, consideration of sound source distance from the listener must be thoroughly investigated and will be considered in detail at this point.

Auditory Depth Perception

Thus far, discussion has focused on the location of auditory sources in reference to their azimuth and elevation. However, environmental sounds emanate from sources located at a distance from the perceiver. This point has been briefly considered in relation to the difference between localization and lateralization but will now be considered in detail. Research on auditory depth perception is sparse, however, like the voluminous literature on depth perception in vision, investigation of egocentric distance estimation in the auditory modality has focused on specific stimulus and contextual cues that give rise to this percept of depth (see Coleman, 1963 for a detailed discussion). The principal cues for detecting static position in depth include (1) intensity cues, (2) spectral absorption cues, and (3) reverberation cues. The nature of each of these characteristics of the sound signal and their relation to the percept of depth will be considered in detail.

The first of these distance cues is conveyed in the overall intensity or sound pressure levels of auditory signals at the listeners point of reference. The increase or decrease in intensity or sound pressure of an acoustic source relative to the listener follows an inverse power function of distance specified by the following equation:

$$\frac{1}{R}\ loss\,/\,gain\,(dB)\ =\ 20\ \log_{10}\frac{R}{R_0} \qquad (6)$$

where R is the distance to a comparison point in depth, ρ , and R_0 is the distance to a referent distance, ρ_0 (Coleman, 1963; Strybel & Perrott, 1984). Therefore, it can be demonstrated that an amplitude decrement of -6 dB corresponds to a doubling of sound source distance and conversely an increase in sound pressure level of +6 dB corresponds to a halving of the egocentric distance of the sound source (Coleman, 1963). It has been shown that listeners can successfully use this intensity information in estimating depth for distances beyond 300 cm (Strybel & Perrott, 1984). At nearer distances, it is evident that these amplitude changes alone are insufficient for adequately estimating egocentric distance. Furthermore, accuracy of sound source distance judgments is a function of the listeners familiarity with the particular sound and performance increases dramatically with practice (Coleman, 1962).

The frequency components of the sound source also have an important influence on the perception of depth. Spectral changes occur as a function of distance because of the different degrees of absorption between high and low frequency spectrum components as the sound wave travels through the air. This differential attenuation of different regions of a complex sound's frequency spectrum leads to detectable changes in the waveform over sufficient distance changes, on the order of 20 to 30 feet, which, in turn, can be used to judge egocentric sound source distances (Coleman, 1963). Consequently, superior distance estimation can be achieved with increasing complexity of the sound source due to the fact that the listener is able to sample the attenuation of sound pressure across a greater number of frequency bands. Of course, this capability is largely dependent on familiarity of the sound source since spectral properties of the sound source may be perturbated without a corresponding change in distance (Coleman, 1963). These perturbations of the sound source spectrum may be the result of absorption as the signal travels through air but can also be indicative of qualitative changes in the signal produced by the sound source. Moreover, the detection of these frequency spectrum changes in cluttered environments is complicated by concomitant distortions introduced by absorption and reflection of the sound wave by environmental structures and materials. However, beside their confounding influence on spectral cues for depth estimation, these distortions of the sound wave by environmental objects can further provide cues to distance assuming familiarity with the acoustic characteristics of these environmental features on behalf of the listener.

Thus, the third significant static cue to auditory localization in depth is reverberation or the reflectance of sounds from environmental obstacles such as walls, ceilings, buildings, and trees. Since each environmental object has different reflectance properties, the ability to specifically describe the manner in which they alter the acoustic spectrum of the sound source is highly intractable (cf., Handel, 1989). Furthermore, since each environmental object differentially alters spectral components of the sound source, the overall effect of the sound environment on distance localization ability can be either beneficial or detrimental. However, in environments in which objects sufficiently attenuate the sound as well as reflect it at decreased amplitude, these reverberations can be used as an absolute cue for depth judgments (Mershon & King, 1975; Mershon, et al., 1989; Mershon & Hutson, 1991). In particular, Mershon, et al., 1989, demonstrated that egocentric distance was underestimated in an acoustically dead room and overestimated in an acoustically live room. Moreover, the presence of background noise served to decrease perceived distance. Finally, distance estimations in highly reverberant environments were shown to become more veridical with increasing exposure. These findings suggest that individuals utilize information regarding the amplitude and time delay between perceptible replications of the sound source reflected from walls, ceilings and other acoustical boundaries. The ability to utilize these cues is mediated by familiarity with the acoustical characteristics of the listening environment.

Thus, research has demonstrated that intensity, spectral changes due to the absorption of air, and spectral changes due to reverberation all contribute to the auditory percept of static egocentric distance. However, the incorporation of these depth cues into virtual acoustic displays has been limited because of their complexity. Such integration requires an acoustic model of the environment being simulated that sufficiently describes the reflectance characteristics of surfaces within this environment. These characteristics must then be incorporated into a spectral transfer function representing perturbations in the acoustic signal created by features of the listening environment. Thus, adequate headphone simulations of free-field listening must include transfer functions representing the acoustic environment as well as the HRTFs that reflect the filtering of the sound source spectrum by the listener's anatomy. Clearly, further research on the role of environmental cues for auditory depth localization is necessary for developing more adequate simulation technology for auditory spatialization.

A more important consideration for incorporating auditory displays in complex environments is the analysis of dynamic cues for perceiving changes in the egocentric depth or distance of auditory sound sources. It could be that dynamic motion in auditory depth is cued by time sampled comparisons of the static depth cues enumerated above. However, other emergent dynamic cues are present and can potentially be used for judgments of velocity and acceleration of approaching and receding sound sources. The first of these dynamic cues for auditory depth perception is the *Doppler effect* (cf., Handel, 1989). The *Doppler effect* refers to the increase in frequency of approaching sound sources and the decrease in frequency of receding sound sources due to compression and rarefaction of the sound wave. Specifically, the *Doppler shift* or change in frequency of a converging or diverging sound source can be expressed as:

$$\Delta f = \frac{u + v}{c} f \qquad (7)$$

where Δf is the resultant change in frequency, v is the speed of the sound source relative to the receiver, u is the speed of the receiver relative to the sound source, c is the speed of sound, and f

is the frequency of the sound source signal (see Kinsler, et al., 1982). The specified change in frequency at the point of reception applies to those cases in which $u << c$ and $v << c$. While the closure rate of the sound source and the receiver must be of significant velocity to yield a perceptible change in the frequency characteristics of a sound, the *Doppler effect* is ubiquitous in the dynamic acoustic environments commonly encountered in our modern technological society. However, systematic empirical investigations of whether and to what extent the *Doppler shift* contributes to an acoustic percept of motion relative to the perceiver are generally lacking and this phenomenon is clearly deserving of further study.

A second dynamic cue for detecting the rate of auditory motion in depth is the acoustic intensity changes generated by sound source convergence or recession (Shaw, McGowan, & Turvey, 1991). It has been mathematically demonstrated that the relative rate of change in intensity, the ratio of the instantaneous sound intensity to its temporal derivative, can be utilized as a dynamic cue to motion in depth and time-to-contact of an acoustic source of relatively small spatial extent. This relative intensity change has been termed *acoustic tau* (τ) because of its functional similarity to the relative rate of size dilation of an approaching visual stimulus termed *tau* (τ) by David Lee (1980) which was discussed in detail earlier. In particular, Shaw, McGowan, & Turvey (1991) demonstrated that the acoustic time-to-contact variable (τ) could be represented by the following equation:

$$\tau = \frac{2I}{{}^{dI}\!/_{dt}} \qquad (8)$$

Where I is the time-averaged intensity of the sound, ${}^{dI}\!/_{dt}$ is the temporal derivative of the intensity indicating its instantaneous rate of change over time, and τ indicates the time-to-contact between the perceiver and the dynamic sound source. Stemming from the inverse-square law describing the relationship between intensity and distance it can be shown that:

$$I = kr^{-2}$$

and

$$\frac{dI}{dr} = -2kr^{-3}$$

therefore,

$$\frac{dI}{dt} = -2kr^{-3}\frac{dr}{dt} = -2kr^{-3}v$$

and

$$\frac{I}{dI/dt} = \frac{kr^{-2}}{-2kr^{-3}v} = \tfrac{1}{2}\tau \tag{9}$$

where *I* is intensity at the point of the perceiver, *k* is a constant indicating sound pressure level of the source, *r* is distance, *v* is velocity, and *t* is time (see Shaw, McGowan & Turvey, 1991 for details). Thus time-to-contact can be acoustically specified by the relative rate of change in the perceived sound intensity. It is posited that this *acoustic tau* (τ) variable can be used perceptually to judge arrival times of sound sources at the listeners location. For this reason, *acoustic tau* (τ) is a critical cue for motor guidance and locomotion in reference to acoustic objects. To date, little empirical evidence has been collected regarding the use of *acoustic tau* (τ) to judge sound motion in depth. However, a study by Schiff and Oldak (1990) demonstrated that blind individuals were as good at judging arrival times of acoustic sources as sighted individuals were at judging arrival times of visual stimuli. Sighted individuals demonstrated poorer performance with auditory stimuli, but this performance improved with increasing presentation times. Thus, it is evident that a cue similar to *acoustic tau* (τ) is employed in estimating velocity and contact times of acoustic objects and utilization of this cue can improve with practice.

Further research investigating the role of dynamic cues to auditory depth is clearly needed. However, rarely is the dynamic percept of location either in depth or in azimuth and elevation using auditory cues devoid of visual information about either the sound source, the surrounding environment, or both. Therefore, detailed discussion of past findings regarding the interactions of vision and audition in the determination of spatial position and spatial dynamics is critical for the understanding of how these spatio-temporal percepts come to specify parameters necessary for motor guidance.

SPATIAL CORRESPONDENCE BETWEEN VISION AND AUDITION

Individuals perceive a unified environment in which objects and features potentially convey spatio-temporal information that is both aurally and visually perceptible. Thus, the nature of how the perceiver comes to equate spatial and temporal information provided by these two distinct modalities is of particular interest. Furthermore, the manner in which information from the two modalities come to facilitate or bias the perceiver's unified dynamic spatial percepts is deserving of detailed inquiry. Consideration of the spatial extent of vision and audition reveals that the visual modality suffers from limitations in the field of view, while the auditory modality is capable of perceiving spatial location and motion through the full 360 degrees in both azimuth and elevation. Therefore, while auditory spatial information is typically of poorer resolution, it can serve to orient the perceiver so that objects can be precisely localized and identified through the visual modality. Recent research that has addressed this spatial orienting role of audition will be discussed in detail. Finally, bimodal facilitation of signal detection and recognition has been extensively documented as a function of the temporal covariation of concomitant visual and auditory signals. These results lead to the conclusion that space and time come to be commonly represented through a pooling of visual and auditory inputs regarding the spatial layout and spatial dynamics of the environment.

A Common Spatial Metric

Foremost in the consideration of auditory and visual spatial interactions is the matter of whether derived auditory and visual spatial relationships are encoded in a common spatial metric and consequently evaluated with a common scale or whether they are encoded and evaluated disjunctively. If auditory and visual spatial percepts are encoded disjunctively, subsequent comparisons between modalities would require translatory processes. On the other hand, if spatial scaling occurs independent of modality, comparisons between auditory and visual spatial cues need not be mediated by translation. This issue is of particular importance because requisite comparisons of auditory and visual spatial position cues in highly dynamic environments must be carried out rapidly without the introduction of large amounts of noise or variance. Auerbach and

Sperling (1974) investigated the plausibility of a common spatial metric representation between vision and audition within a signal detection theory (SDT) framework.

Auerbach and Sperling (1974) reasoned that any translatory process between disjunct spatial representations would be revealed through an increase in response variance based upon the assumption that all human information-processing systems are inherently prone to noise and equivocation of their input signals. Based upon this assumption, Auerbach and Sperling (1974) posited that response variance in a two alternative forced choice (2AFC) spatial task, would demonstrate higher variance levels when the comparison was between modalities as compared to when the comparison was within a single modality, if and only if, a translatory processes had mediated the cross-modal comparison. Moreover, in signal detection terms, the sensitivity (d') for detecting apparent position would be reduced in cross-modal conditions if translational processes occurred. Same-different responses to spatial position of temporally separated presentations of two lights, two spatially localized tones, or combinations of lights and tones revealed that sensitivity (d') of spatial estimations for light-tone combinations was sufficiently accounted for by the variance of unimodal responding. That is, the sensitivity (d') measures empirically observed corresponded to theoretical predictions assuming a common spatial metric and were not significantly decremented by a hypothesized translational process as would be predicted by a disjunctive account of auditory and visual spatial encoding.

The conclusions to be drawn regarding a common metric for auditory and visual spatial comparisons suffer from the fact that they draw upon empirical research that sought to verify the null hypothesis. Nonetheless, the implications of these findings are critical for understanding performance in multimodal spatial tasks. If one is willing to accept the common spatial metric hypothesis in light of the fact that it is not empirically falsifiable, functional accounts of speed-accuracy tradeoffs in motor tasks directed at bimodal stimuli can be sufficiently described perceptually by the static and dynamic spatial resolution of the auditory and visual systems. That is, models of performance need not rely upon functional descriptions of translational processes mediating between auditory and visual spatial percepts. However, this does not imply that auditory and visual spatial percepts contribute to the overall common spatial representation of an object independently. In the following section, the manner in which visual and auditory cues to spatial location interact and the facilitory and biasing effects that arise form these cross-modal spatial cue interactions will be examined in detail.

Visual Cues for Auditory Spatialization

Since auditory and visual spatial comparisons apparently can be made without translatory processing, spatialization abilities in the auditory modality can be augmented by visual cues. For example, auditory spatial information in sighted individuals is facilitated when responses are made with the eyes open and when textural cues can be perceived (Warren, 1970; Platt & Warren, 1972). In particular, Platt and Warren (1972) demonstrated that localization of an auditory tone, measured by both mean absolute error and variance of hand pointing responses, was best when sufficient light was available to view a textured background and eye movements were permitted during localization. Localization ability was significantly poorer in conditions in which subjects were instructed to maintain optical fixation during localization and in conditions where visual information was not available. In a second experiment, the addition of a condition in which eye movements were permitted during auditory localization in a completely darkened environment indicated superior localization performance as compared to conditions in which the subjects maintained visual fixation in a lighted environment. Based on these findings, Platt and Warren (1972) concluded that the facilitation of auditory spatial discriminations provided by vision was the result of an interaction between eye movements and the perception of textural features. They further reasoned that this effect on auditory spatialization was resultant from the contribution of visual spatial cues whose accuracy is improved by feedback based upon comparisons of optomotor afference or efference and the resulting optical flow of textural features.

Given the requisite of a complex symbiosis between optomotor and optical parameters for visual cueing of auditory localization, it is reasonable to assume that such facilitory effects would emerge over the course of perceptual development. Moreover, the experiential role of vision as the dominant modality for conveying information about spatial layout in sighted individuals may be revealed through a developmental emergence of visual facilitation effects for spatial tasks. In a series of experiments, Warren (1970) demonstrated that the superiority of auditory localization responses made in the presence of visual cues was only evident in adults, and prior to adolescence, auditory localization was found to be superior in the absence of visual information rather than in its presence. Thus, the ability to make cross-modal comparisons of spatial information appears to be learned through interaction with spatial objects and events that exhibit cues to localization perceptible by both vision and audition. This developmental pattern provides further insight into the nature of the common spatial representation between vision and audition. Since this spatial encoding appears to be learned, it can be presumed to be biased in favor of information that dominates the spatial layout. The greater spatial acuity of the visual system coupled with the fact that a significant proportion of environmental information regarding spatialization is conveyed visually gives rise to the dominance of the visual modality in spatial perception. This dominance can be evidenced through perceptual biases favoring vision that occur when discrepancies or conflicts arise between auditory and visual spatial cues.

In the presence of discrepant spatial information from the auditory and visual modalities, the position of the referent object is biased by weightings of the auditory and visual cues available in the given context so as to maintain unity within the multimodal spatial percept (cf., Welsh & Warren, 1980; Welsh & Warren, 1986). The relative superiority of the visual system in terms of acuity typically leads to greater weighting of visual spatial cues. This superiority of visual cues leads to the empirically observed bias of spatial position estimations in favor of the visual object under situations of cross-modal spatial cue conflict. Consequently, this visual superiority has been termed the *visual capture* phenomenon or simply *visual dominance* (see Welsh & Warren, 1980; Welsh & Warren, 1986). This perceptual bias has also been referred to as the *ventriloquism effect* because it is presumed to be the vehicle that allows ventriloquists to project their voices through the mouths of their dummies. Functionally, an assumption of unity between visual and auditory percepts on behalf of the perceiver in these situations is presumed to lead to a process of perceptual normalization to resolve intersensory discrepancies. The emergence of biases in the perceptual normalization process can be accounted for by three complementary theoretical accounts: (1) modality precision, (2) directed attention, and (3) modality appropriateness (Welsh & Warren, 1980; Welsh & Warren, 1986). Modality precision suggests that the modality with greater spatial resolution will receive higher weighting in estimations regarding location. The directed-attention hypothesis suggests that modalities are weighted by the distribution of attention allocated between them. Consequently, manipulations to direct attention toward or away from a particular modality can serve to increase or decrease the biasing effects of that modality.

Typically, attention in spatial tasks is directed toward visual cues because of the prevalence of vision in conveying information regarding spatial layout. Finally, the modality appropriateness hypothesis suggests that modalities will receive biasing weights in multimodal estimations to account for their particular adaptive evolution for perceiving differences along the dimension being evaluated. Thus, while audition has specifically evolved to be highly sensitive to temporal patterning, the visual modality is specifically adapted to perform tasks involving spatial judgments. Of course, this modality appropriateness is reflected through the relative spatial precision of vision and audition. In any case, the ubiquitous bias toward vision in spatial judgments can be attributed to the appropriateness of vision in such tasks due to its highly acute spatial resolution. Furthermore, this bias can be mediated by situation specific parameters such as information quality between modalities and attentional focus. In any case, its antecedents are clearly historical in nature and are learned through years of interaction with an environment that is predominated by visual cues to spatial location. Consequently, the visual dominance effect, like the visual facilitation of auditory localization judgments discussed above, is not witnessed in young subjects (Welsh & Warren, 1980). The impact of this intersensory bias toward the visual stimulus in motor task performance is a matter that remains to be addressed in a systematic empirical research endeavor. However, despite the dominance of the visual modality both in terms of its acuity and its predominance in the environment, visual guidance in motor coordination is limited by a restricted field of view and an even smaller focal region in human perceivers. Thus, spatial information from the auditory modality, which is not as spatially constrained, can potentially be used to cue location of important visual features requiring attention. Consequently, the influence of auditory spatial information on visual task performance is deserving of detailed consideration.

Auditory Cues for Visual Spatialization

Recent research has examined the benefit of providing auditory spatial cues in visual search tasks (Perrott, et al., 1990; Perrott, et al., 1991). These studies demonstrated that "the auditory spatial channel has a significant role in regulating visual gaze" (Perrott, et al., 1990, p. 214). In particular, using two alternative forced choice (2AFC) search tasks, it was demonstrated that reaction time was significantly reduced when a spatialized auditory cue was presented during search trials as compared to trials in which localization and identification were carried out through vision alone. This auditory cue provided information regarding the location of the visual target but gave no further information regarding the nature of that target. In this sense, the auditory information had no bearing on task accuracy. Nonetheless, the presentation of this auditory information did significantly reduce the amount of time required to complete a trial without influencing error rates. The benefit of the simultaneous auditory spatial cue presentation was most pronounced when the visual target was in the periphery, outside the field of view. However, benefits were also evident when the visual target was within the field of view and even when the visual target was at the fixation point set prior to trial initiation. This final result of auditory aiding for visual targets at the fixation point is particularly striking. It could be the result of a reduction in uncertainty provided by the auditory spatial cue that prevented rapid visual scanning away from the fixation point. This effect could also have been mediated by the superior alerting of attentional mechanisms supported by the auditory modality as compared to the visual modality (cf., Posner, 1986). However, results are more pronounced at more peripheral locations where spatial information was particularly critical for task performance. At these peripheral locations, the decrease in reaction time in the 2AFC tasks under auditory cueing conditions was dramatic, in excess of 300 ms (Perrott, et al., 1991). Furthermore, these performance gains obtained from the addition of auditory spatial information were most pronounced when a large number of distractors were present in the visual task. Thus, auditory aiding appeared to be most beneficial when the visual display was highly cluttered thus rendering the search task more difficult.

Further research investigating the application of localized auditory cues for visual detection in an applied task domain was conducted by Begault (1993). Begault (1993) investigated the performance benefits of providing spatialized auditory information in an aviation traffic collision avoidance system (TCAS). Twelve groups of two person commercial airline crews conducted simulated flight missions in visual conditions. Half of the crews completed the flight with the benefit of three-dimensional localized auditory traffic advisories presented over headphones using digitally filtered HRTFs. The other half of the crews completed their missions using a TCAS system that presented auditory traffic alerts through their headphones monaurally. Among missions completed using the localized auditory TCAS system, auditory traffic advisories coincided spatially and temporally with visual presentations of 24 target aircraft images that posed a potential collision threat. Among missions completed using the non-localized auditory TCAS system, auditory advisories temporally coincided with visual presentations of the 24 target aircraft but conveyed no position information. Results demonstrated that flight crews using the localized auditory TCAS system visually acquired the target aircraft images significantly faster than their counterparts who completed the simulated flights using the non-localized auditory TCAS display. A mean visual acquisition time improvement of 2.2 seconds was demonstrated among flight crews flying with the localized auditory TCAS display, clearly indicating a significant advantage over conventional non-localized auditory TCAS displays and thereby allowing greater response times to maneuver the aircraft in time critical scenarios. Similar ongoing research efforts within the United States Air Force Armstrong Laboratory has focused on the potential application of 3-D audio systems for target acquisition (see McKinley, et al., 1994). Results from laboratory and in-flight demonstrations of these systems have indicated that 3-D auditory cuing provides combat aviators with increased situational awareness and a concomitant decrease in subjective workload. Pilots in these studies demonstrated decreased visual target acquisition times when 3-D auditory cues were provided as compared to conditions in which target searches were made using vision alone.

To further assess whether both position and velocity information conveyed aurally could be used as preview for visual objects prior to their entry into the field of view, I completed a study in which subjects completed a visual target aiming task with varying duration and quality of the auditory preview (see Elias, 1994). During this study, subjects completed a target aiming task in which they made discrete button press responses to fire a computer graphic projectile at a moving target image. After practicing this task using only visual cues, dynamic auditory preview created by moving a speaker located on a linear slide in synchrony with the visual target image was introduced: This dynamic auditory preview was presented to subjects at various distances beyond the bounds of the visual display. Results indicated that when the target moved too rapidly to make an adequate firing response visually, auditory preview information improved performance. Furthermore, performance improved with increasing auditory preview distance, indicating that with more time to sample the auditory preview information, more accurate estimations of the velocity and time of arrival of the visual target at the optimal firing point could be determined and consequently performance improved. However, there was a diminishing marginal gain in performance with increasing auditory preview distance, indicating that once sufficient auditory preview was provided, the further addition of auditory preview information had little impact on task performance. In subsequent testing, the quality of the auditory preview information was altered by offsetting the sound source from the visual target position by some distance, or by having the sound source travel faster or slower than the visual target. In position misalignment conditions where the sound source lagged behind the visual target, higher error magnitudes were observed. However, when the auditory display preceded the visual target, performance actually improved, presumably by compensating for inherent perceptual-motor delays in visual responding. Similarly, in velocity mismatch conditions, responses toward fast moving targets improved when a relatively faster sound source was previewed. However, responses were disrupted when a slower sound source was previewed. On the contrary, responses toward slow moving targets improved when a relatively slower sound source was previewed but were disrupted when a faster sound source was previewed. This observation can be attributed to the phenomenon of *priming* (c.f., Posner, 1986) whereby the preview of a relatively slower or faster target biases responding in a manner that compensates for inherent anticipatory responses and perceptual-motor delays thereby reducing errors attributable to responding too hastily to a slow moving target and not responding quick enough when the target is traveling at a high velocity.

These dramatic demonstrations of performance benefits in visual search and visual target aiming through auditory spatial cueing lead one to ponder whether similar cues could be utilized to aid performance in making temporally and spatially precise responses in motor guidance tasks. Clearly, research has indicated that the spatial covariation of auditory and visual stimuli is one significant contributing factor mediating the facilitory effects and biases that occur between modalities. A second critical factor influencing factor contributing to the utility of bimodal signals is the temporal covariation of auditory and visual stimuli. Indeed, the utility of bimodal signals is further influenced by the temporal covariation of the auditory and visual stimuli as well as their perceptual integrality. In the following discussion, the bimodal facilitation of task performance will be discussed as a function of the temporal relationships that exist between auditory and visual signals.

Bimodal Facilitation

·Thus far, discussion has been devoted to spatial comparisons between vision and audition. However, the organization of bimodal percepts is best defined in terms of its spatio-temporal properties. That is, bimodal facilitory effects are not only influenced by the spatial correlation of signals in each modality, but are further influenced by the temporal synchrony between these signals. O'Leary and Rhodes (1984) demonstrated that these spatio-temporal characteristics of auditory and visual signals can serve to bias perception in the other modality. In particular, segregation of objects in time sampled displays is influenced by stimulus onset asynchrony (SOA) or interstimulus interval (ISI) in both visual and auditory displays. In visual displays, long SOAs between frames in which a stimulus is displaced produces a percept of a single object in motion (solid lines in Figure 19). However, short SOAs produce a percept of two distinct objects in motion over a more restricted spatial range (dashed lines in Figure 19). Based upon the temporal patterning of auditory and visual displays, the ability of object segregation in one modality to influence spatio-temporal object segregation in the other modality was demonstrated. In particular, SOA thresholds for the perception of one or two objects in a target modality were determined as a function of the percept presented in the secondary concomitant modality. Auditory and visual stream segregation occurred in both modalities at higher SOAs when information from the other modality gave rise to a percept of two distinct objects. These results indicated a cross-modal bias in spatio-temporal perceptual organization. These results further suggest that auditory and visual modality interactions are formulated through their temporal covariance as well as their spatial covariance.

In fact, the vast majority of research addressing cross-modal interactions between vision and audition has focused upon the temporal synchrony of bimodal stimulus presentations. The redundancy gain attainable when perceivers are presented with temporally covarying auditory and visual signals has been extensively documented in the literature (see, e.g., Brown & Hopkins, 1967; Fidell, 1970; Loveless, Brebner & Hamilton, 1970; Doll & Hanna, 1989). Specifically, greater sensitivity to signal presentations has been demonstrated with bimodal as opposed to unimodal signal occurrence. For example, Fidell (1970) found that signal detection of a sinusoid can be improved when the signal occurs in both the visual and the auditory modality rather than in either of these two modalities alone. These bimodal effects are presumed to result from either reduction in the uncertainty of signal occurrence, reduction in noise, or some combination thereof. Reduction in uncertainty can result from additivity between the two modalities. That is, the additive probability of detecting two stochastic signals is at least as great as and potentially greater than the probability of detecting either of the signals alone. Noise reduction, on the other hand, can only be accomplished by a pooling of information from modalities into a common representation of auditory and visual inputs. That is, the randomness of noise within a modality suggests that the temporal covariation of noise elements between modalities will have chance probability. Consequently, a common temporal representation of bimodal information could be structured in which event likelihood is interactively determined based upon the commonality of information across modalities. Thus, the signal-to-noise ratio of bimodal events can be increased through additive and/or interactive pooling of auditory and visual signals (cf., Fidell, 1970; Loveless, Brebner, Hamilton, 1970).

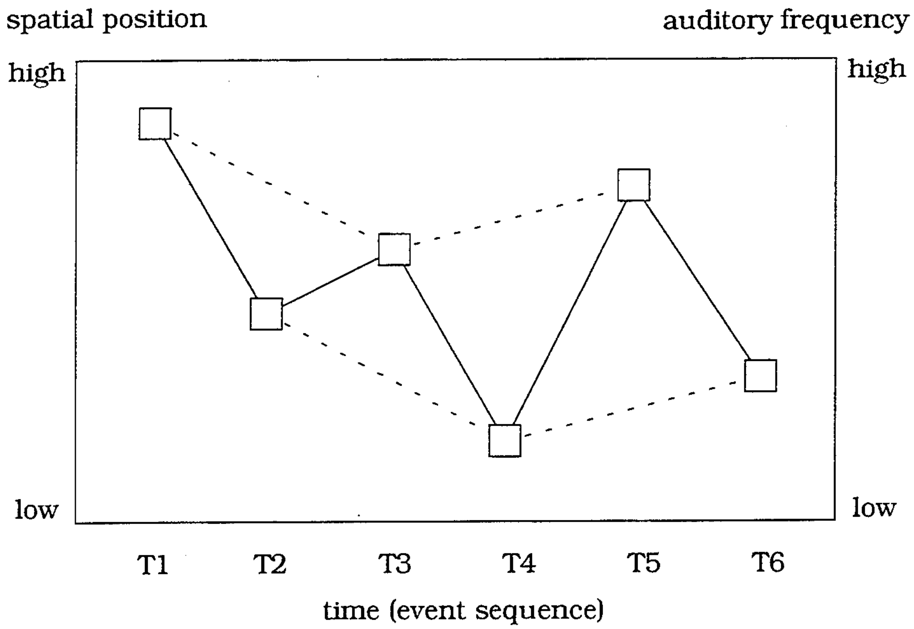

Figure 19. Stimuli used in O'Leary and Rhodes (1984). Spatial y-axis position and auditory frequency were systematically varied. Based upon SOA either a single object percept (solid lines) or a percept of two distinct objects (dashed lines) could occur.

This notion of multimodal pooling can be extended to the spatial domain. That is, just as noise can be attenuated by interactive comparisons of temporal covariation between modalities, noise attenuation can be mediated by comparisons of spatial position across modalities. Doll and Hanna (1989) attempted to address such a possibility by presenting lateralized auditory tones and non-lateralized auditory tones during bimodal signal presentation. Unfortunately, no effect of spatial compatibility was found in their study. However, as they correctly asserted, the lateralized auditory tones employed were insufficient for providing spatial information. Therefore, enhanced bimodal compatibility through localized auditory cues remains a topic worthy of empirical investigation. The evidence presented thus far in relation to auditory and visual spatial interactions suggests that such bimodal spatial facilitation effects should be demonstrable and should interact with the temporal covariation of bimodal signals. Thus, auditory and visual information appears to be conveyed within a common spatio-temporal representation upon which signal detection discriminations are made. Moreover, this common spatio-temporal representation provides the necessary parameters for motor programs that signal effectors to operate on the environment (cf., Pew & Rosenbaum, 1988). That is, the requisite parameters for successful guidance of motor responses lie in the spatio-temporal representation of the environment. A unified spatio-temporal representation based upon weighted additive and interactive contributions of signals from the auditory and visual modalities can convey such information to motor effectors in a functionally effective and parsimonious manner. Consequently, this spatio-temporal representation of the environment can serve to control and guide complex motor activities such as rapid aimed movements and continuous manual control manipulations. The characteristics of these motor actions will consequently serve as the focus of further discussion regarding this hypothesized common spatio-temporal representation.

CONTROL OF MOTOR ACTIONS

In the preceding discussion, it was demonstrated that computational processes involving visual and auditory perception can operate synergistically to produce a unified representation of the spatial layout and spatial dynamics of the environment. The parameters specified in this spatio-temporal representation provide the requisite information for motor guidance programs. In the following discussion, the role of these motor programs in directing complex behaviors such as rapid aimed movements, continuous manual control, and egomotion will be considered in detail. The critical determination of speed and accuracy in performing these tasks relies upon the specification of spatio-temporal characteristics of the environment as well as the specification of mass properties of the organisms structural anatomy and objects of motoric manipulation. The spatio-temporal specification of the motor action is presumed to arise from the multimodal representation of the dynamic environment, whereas mass properties are assumed to be specified though learning, prior history of the organism, and propioceptive feedback.

Rapid Aimed Movements

Rapid ballistic movements of the extremities are a particular topic of interest because they comprise the building blocks of more complex motor activities. The nature of how these actions are planned and executed is inherently linked to perceptual information regarding spatial layout and spatial dynamics. That is, the parameters requisite for successful completion of these movements must specify the timing and coordination of anatomical structures in reference to the spatial relationships between the actor and the layout of the environment. However, even the simplest of observable motions can be the result of a complex interaction between a large number of effector muscles and joints. Therefore, the specification of a motor action must reflect the manner in which these skeleto-muscular structures will interact and respond synergistically as a single functional unit (Pew & Rosenbaum, 1988). Research indicates that at the highest level these plans for motor action are specified in terms of the spatio-temporal relationships between environmental features and the anatomical structures that directly contact them. For example, analyses of motion trajectories indicate that the planning of hand movements is specified in terms of hand-space, specifying the relative position of the hand, rather than in terms of a joint-space specification of the relative positions of the various joints required for movement. The benefit of such a plan is reflected in its economy in that it minimizes the degrees of freedom that must be parameterized. In so doing, this planning occurs at a high level of abstraction in terms of motor functioning (see Pew & Rosenbaum, 1988). Nonetheless, the plan is highly specified in terms of the spatio-temporal layout of the environment. That is, it potentially contains all requisite perceptual information regarding spatial layout and spatial dynamics. In this manner, ballistic motor actions can be initiated with adequate specification of spatio-temporal parameters characterizing the dynamic environment in which these actions will occur.

The action plans formulated from these spatio-temporal parameters serve to guide motor responses. Schmidt, et al. (1979) posit that the invariant features of such motor responses fall into two major categories: (1) those that specify the phasing or temporal relationships between components of a response, and (2) those that specify the relative forces required for execution. These invariants can thus be parameterized into motor programs to specify the coordination of motor effectors that produce a desired response goal. At the highest level, these invariants must receive input regarding the spatio-temporal characteristics of the environment in order to construct the response goal. Since phasing and force parameters would similarly be defined spatio-temporally, they can be realized by computational solutions that incorporate perceptually derived space and time parameters as well as mass and force parameters obtained from learning and the prior history of the organism. These mass and force parameters can be equated with Schmidt's (1975) concept of a *motor schema.* Schmidt (1975) suggests that these schemas are comprised of four properties of goal directed movements: (1) initial situational conditions, (2) response specifications for the motor program, (3) sensory consequences of the response to be produced, and (4) the outcome of the movement. The first of these components, the initial conditions, reflects the spatio-temporal representation of the environment in which the task is being performed. Second, specifications for the motor program are conveyed through force-time parameters. Finally, sensory consequences serve to convey information regarding success or failure and degree of error. Consequently, a schema representing triggering conditions and force-time constraints can be built up through repeated practice and experience. That is, when knowledge of results regarding the spatio-temporal outcomes of a movement relative to objects in the environment are available, they serve to reinforce the application of force parameters that lead to successful motor response outcomes. In fact, the relationship between knowledge of results and reinforcement is striking. Like schedules of reinforcement, knowledge of results provided visually or aurally produce the most accurate learning of response force requirements when they are presented in temporal proximity to the response occurrence and when they are presented on an intermittent schedule (see Salmoni, Schmidt, & Walter, 1984; Schmidt, et al., 1989). Furthermore, the amount of information and the accuracy of information conveyed through in these transmittals of knowledge of results directly impacts the degree of improvement in performance. Specifically, presentation of information regarding the absolute disparity between actual and desired outcomes produces superior learning as compared to presentations of relative outcome disparities over the course of learning a motor skill. Hence, force parameters associated with rapid aimed movements can be learned in a general manner and functionally operate to increase the speed and decrease the variability of responsive motor actions over the course of skill acquisition (cf., Adams, 1987). Thus, during early stages of determining relationships between muscular force, spatial movement amplitude, and time, knowledge of results are critical for reducing the variability of responses. Adams (see Adams, 1987 for a review) consequently described the stage of early learning that he termed the *verbal-motor* stage as being closed-loop in nature and contingent upon knowledge of results for subsequent learning of the force constraints imposed on a motor task. Adams further suggested that the reliance on knowledge of results diminishes with practice and consequently, after the motor skill has been well learned, individuals enter what Adams refers to as the *motor* stage of skill development. In this late stage of skill acquisition, improvements can continue without knowledge of results because at this point proprioceptive feedback is presumably capable of conveying information regarding differences between the desired and actual amounts of force applied in the execution of motor programs and can serve to iteratively reinforce movements whose force components fall within an increasingly restrictive range of tolerance (see Adams, 1987; Mazur, 1990 for overviews of motor learning theories).

Thus, a practiced motor activity is based upon initial planning in terms of direct contact relationships between objects and the actor, the programming of motoric movements constrained by learned force-time parameters specified through prior experience and current spatio-temporal constraints, and the execution of these programs by skeleto-muscular effectors. These functions are reflected in the five phases of rapid motor actions defined by Walker, Meyer, & Smelcer (1993): (1) goal formation, (2) response selection, (3) motor programming, (4) movement execution, and (5) movement verification. Accordingly, goal formation and movement verification serve a functional role in coordinating perception and action. That is, goal formation can inherently be defined in terms of differences between actual and desired states in the spatial layout and spatial dynamics of the environment and can be conveyed through the affordances of objects (cf., Gibson, 1979). Moreover, movement verification can serve to describe the relationship between the spatio-temporal characteristics of the environment after movement execution and the desired state of the spatial layout thereby providing perceptual assessment of feedback and knowledge of results. In this manner, precision movements can be learned and performed through computational processes that describe the spatio-temporal relationships and mass properties of the actor and the environment. These precision movements form the foundation of complex motor actions, and consequently the specification of parameters for these movements will be considered in greater detail.

In cases of discrete goal directed movements toward stationary objects, a lawful relationship between spatial parameters and motor performance has been repeatedly demonstrated. This lawful relationship, Fitts' Law, posits a speed-accuracy tradeoff function which can be expressed by the following equation:

$$MT = a + b\log_2 \frac{2A}{W} \qquad (10)$$

where *MT* is movement time, *A* is the amplitude or distance to the target object, *W* is the target width, and *a* and *b* are constants (see, e.g., Schmidt, et al., 1979; Meyer, et al., 1988). The *Index of Difficulty* of a movement is defined within this equation as *2A/W* and has repeatedly been shown to be a sufficient parameter for describing movement performance. However, more recent research has suggested that the relationship between this Index of Difficulty and movement performance is better described as a power law function:

$$MT = a + b\left(\frac{A}{W}\right)^{p} \tag{11}$$

where a and b are non-negative constants and $0 < p \le 1$. Values for p have been empirically determined to be in the range of 0.25 and 0.5 (see Walker, Meyer, & Smelcer, 1993). The specific characteristics of this power law relationship are presumed to be mediated by the microstructure of the movement as well as by task constraints (Meyer, et al., 1988).

First, the microstructure of movement velocities and trajectories is assumed to arise from inherent noise in the transmittal of neuromotor signals. Early theories held that this noise was functionally attenuated through successive approximations toward the movement goal through deterministic iterative corrections of the movement's trajectory and force parameters (see Meyer, et al., 1988 for a through discussion). However, more recent theories have demonstrated that the speed of many of these movements renders them incapable of receiving sufficient sensory feedback to guide successive iterations. Consequently, stochastic variations in force parameters of rapid aimed movements are assumed to be corrected through estimation and correction of motor programming error through a minimum number of submovements that serve to optimize movement time (Meyer, et al.; 1988; Meyer, et al., 1990). Thus, this optimized submovement model of rapid movements serves to integrate the notion of iterative corrections of force parameters and the inherent error or variability in force impulses due to psychomotor noise (cf., Walker, Meyer, & Smelcer, 1993). Consequently, it suggests that the speed-accuracy relationship specified in Equation 10 is inherently dependent on the quality of the initial specification of spatio-temporal response parameters as well as the degree of stochastic variability displayed centrally and peripherally by the specific psychomotor pathway.

The nature of this speed-accuracy tradeoff is further influenced by movement requirements or task specific parameters. First, performance relationships are dependent upon the spatial constraints and the temporal constraints of the task. That is, an empirical distinction must be made between spatially constrained tasks, such as those considered above, and temporally constrained tasks (cf., Meyer, et al., 1988). While subjects performing spatial accuracy tasks are instructed to move as quickly and accurately as possible, subjects performing temporal accuracy tasks are instructed to move to the target in a manner such that *MT* = t (where *MT* is movement time, and t is a constant). In these temporal accuracy tasks, subjects variability in movement distance can be specified by the following equation:

$$s^2 = \left(a + b\frac{A}{MT}\right)^2 \tag{12}$$

Where s^2 = movement variance, *A*= movement amplitude or distance, *MT* = empirical movement time, and *a* and *b* are positive constants. Thus, movement velocity, *A*/*MT*, is a sufficient parameter for predicting movement variability. Consequently, temporal constraints serve to specify required movement velocity which directly impacts movement endpoint variability. Additionally, this role of velocity in determining motor performance variability is evidenced through the effects of control relationships in indirect manipulation tasks. That is, the nature of a control system has a significant impact on the variability and speed-accuracy relationships in movements made with control devices. In particular, Jagacinski, et al. (1980) demonstrated that motor performance using first order velocity controls exhibit steeper slopes in speed-accuracy tradeoff functions as compared to zero order position control systems. Moreover, the control gain in such systems interacts with control order to influence task difficulty and the functional relationships between spatio-temporal parameters and task performance. Thus, evidence from experimental manipulations of temporal constraints and motor control order demonstrate the importance of spatio-temporal control characteristics in specifying the nature of responsive motor output. This is significant since real-world motor tasks that are both spatially and temporally constrained are ubiquitous.

Furthermore, unlike the experimental tasks considered thus far, real-world motor tasks are rarely constrained to one axis of movement. Typical movements can be made at least along two dimensions and frequently are made along all three spatial axes. Nonetheless, research has indicated that these controlling relationships between spatial and temporal parameters described above can be extended to control movements in two and three dimensions (see, e.g., Jagacinski &

Monk, 1985; Walker, Meyer, & Smelcer, in press). However, in reference to these multiple axes of movement, coordination of muscular activity must be extended to account for variability in movement space over multiple dimensions. Empirical observation of these multi-axis trajectories indicate that they fail to conform to a simple Euclidean computation of movement distance. Nor is movement along these multiple axes parameterized completely independently as would be suggested by a city block model of movement space. Rather, these spatial parameters appear to be partially integrated and the degree of this integration is mediated by the asymmetric coding of force parameters along multiple dimensions (cf., Jagacinski & Monk, 1985). This ability to integrate spatial parameters along multiple dimensions may be a function of past experience with dimensional mass and force characteristics of particular effectors along multi-axis trajectories. For example, Jagacinski & Monk (1985) found that two-dimensional aimed movements of the head were more integrated with respect to movement axes than two-dimensional aimed movements of the hand. In any case, the specification of spatial relationships along multiple dimensions leads to lawful functional relationships between the speed and accuracy of control movement as demonstrated in more restrictive motion along a single axis.

Thus far, discussion has focused on rapid movements of the extremities, however, rapid eye movements or saccades have been shown to conform to these functional relationships regarding spatio-temporal characteristics of the planned movement trajectory and the speed and accuracy of performance (see Abrams, Meyer, & Kornblum, 1989; Abrams, 1992). These saccades consequently have important roles in directing guidance of other motor actions that directly manipulate target objects. However, as was already discussed, the visual modality suffers from a limited field of view and consequently, saccades must interact with head and body movements to spatially localize highly eccentric targets. Earlier, research on visual search suggested that this interactive guidance process can be parameterized by spatial information presented in the auditory modality (viz., Perrott, et al., 1990; Perrott, et al., 1991). These findings demonstrated beneficial effects in visual search times when auditory spatial cues were provided. In my doctoral research (Elias, 1994), I demonstrated that dynamic auditory preview of a moving target significantly improved visual target aiming when the sound source and the visual target were aligned and when the auditory cue was misaligned to compensate for inherent perceptual-motor delays. Furthermore, misalignments between the sound source and the visual target that caused delays in anticipated arrival of the target yielded poorer task accuracy. While these findings provide important insight into the nature of bimodal response parameterization, specific research addressing rapid aimed movement characteristics of responses to auditory or bimodal targets has not been a specific topic of study to date. Evidence of lawful speed-accuracy performance relationships such as those described above would provide further evidence for a similar means of representing auditory and visual spatial information.

The guidance of such orienting responses directed to highly eccentric spatial locations specified by auditory targets further highlights the role of large scale motor coordination. That is, a complex symbiosis must exist between motor programs specifying head and eye movement in order to optimally approach such a target. In addition to motor unit specific force parameters, these motor programs must convey precise timing information regarding the relative initiations of head and eye movements. Evidence from empirical studies of complex motor movement sequencing suggests that the timing of these movements is centrally specified and controlled hierarchically (Pew & Rosenbaum, 1988; cf., Pew, 1974). Specifically, evidence has demonstrated that complex control requiring movement of multiple extremities involves temporal coordination that is centrally parameterized. Consequently, the timing of temporally confined discrete motor actions cannot be carried out independently and must be synchronized. Moreover, parameterization of these temporally coupled actions can be economized by consideration of response similarities thereby forming hierarchical chains of movement coordination. For example, in analyses of sequential finger tapping, inter-response delays demonstrate temporal groupings between movements that can be similarly parameterized (see Pew & Rosenbaum, 1988). Specifically, sequential tapping movements of the identical fingers on the opposite hand can be produced more rapidly than sequential tapping of different fingers. This suggests that parameterization is economized by a realization that the two temporally associated actions involve effectors with common mass properties thereby allowing force parameters to be specified in a unitary fashion. Similarly, Pew (1974) described the strategic extraction of rhythmic regularities in forcing functions to allow for the temporal synchronization of movement plans for continuous tracking. Thus, through central time-keeping and hierarchical specification of temporally correlated actions, complex coordination can be achieved in highly integrated motor task performance. The manner in which this coordination could function in the orientation and direction of gaze has already been suggested. This coordination of spatial and temporal response characteristics across multiple effectors can also play a significant role in tasks where target objects are continually displaced. The ability to track such objects both through the direction of visual gaze and complex motor coordination of the extremities is critical in order to continually receive feedback and update motor program parameters to reflect the dynamic changes in the task environment (Abrams, 1992). Such perceptual-motor interdependencies have been extensively studied in reference to continuous manual control in tracking tasks which will now be considered in detail.

Continuous Manual Control

In previous discussion, the precise functional relationships between parameters of the spatial layout and aimed movements were explored. Additionally, the means through which sequential motor movements subsumed under a single action plan are coordinated were considered briefly. Such coordinative sequencing of motor actions is critical for responding in highly dynamic environments where the spatio-temporal characteristics of referent objects are in constant flux. In response to this dynamic flux, sequences of motor actions must additionally be coordinated with sensory feedback regarding the spatio-temporal changes in the multimodal information array. Moreover, this coordination must perform time-based corrections to update information regarding changes in spatial layout and account for stochastic sensori-motor delays in the course of information processing. In engineering domains, this coordination is further complicated by noise and equivocation inherent in control systems. Human operators performing tracking tasks in such domains can be described in terms of the manner in which they continually emit coordinated sequences of actions to minimize disturbances in system parameters under their control. These action inputs are mediated by the dynamics of the engineered system and are represented in a system display where their effects on the status of the system are conveyed to the user (cf. Wickens, 1984). Examples of tracking are ubiquitous in engineered transportation systems such as automobiles, ships, and airplanes, but are also evidenced in domains such as continuous process control and telerobotics in manufacturing. In transportation systems, tracking is directly related to changes in the spatial layout associated with movement of the self or movement of referent objects. In other scenarios, the referent display may directly convey spatio-temporal changes or may convey information that temporally varies but is non-spatial in character. For example, in continuous process control, parameters may be conveyed graphically or spatially to aid visualization of their dynamics but nonetheless are inherently non-spatial. In any case, motor planning and sequencing to achieve system control in these domains can be similarly parameterized through the spatio-temporal dynamics of display properties and are accounted for in control theoretic approaches to sensori-motor coordination.

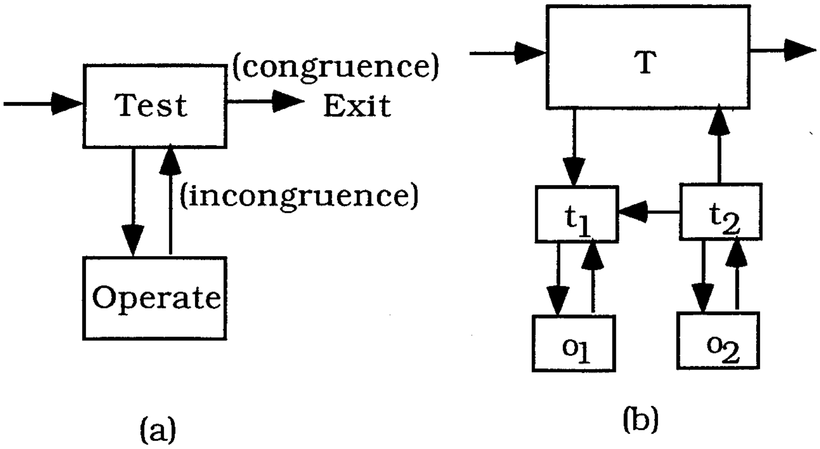

Figure 20. The iterative scheme of motor planning and control defined by Test-Operate-Test-Exit (*TOTE*) units (a) that can be hierarchically arranged (b) to carry out motor actions that serve to minimize disturbances or differences between an actual and desired state (after Miller, Galanter, & Pribram, 1960).

Early models of motor planning described this sensori-motor coordination in terms of hierarchical chains of servo-control feedback loops. For example, Miller, Galanter, and Pribram (1960) described an iterative feedback loop as the functional unit of a motor plan. They posited that this feedback loop is comprised of an iterative sequence of functional steps in which the operator tests the relationship between actual and desired states, operates to minimize disturbances, and final exits the control loop when congruence exists between the actual and desired state of the system. Miller, Galanter, and Pribram (1960) thus described this iteration as a Test-Operate-Test-Exit (*TOTE*) unit of motor responding (see Figure 20a). In the execution of complex motor plans, hierarchical and chained configurations of these *TOTE* units, such as that shown in Figure 20b, can subserve the realization of a motor response goal. Earlier, it was shown how these hierarchical chained arrangements can serve to coordinate and economize the spatio-temporal parameterization of discrete motor responses. The notion of a *TOTE* unit is particularly applicable in these applications, but it also serves to describe the iterative coordination of sensory feedback and motor responding in highly dynamic situations. That is, in a dynamic system where disturbances in spatio-temporal parameters create incongruence between actual and desired states, operator performance can be characterized through hierarchical arrangements of *TOTE* units defined through task analytic techniques. However, the continuous iterative processing of minimizing incongruities in highly dynamic systems is further complicated by temporal delays and sensori-motor noise introduced at all levels of information processing that is not accounted for in Miller, Galanter, and Pribram's (1960) model of motor processing. Therefore, models of continuous manual control behavior have been conceptualized to account for the predictive and corrective nature of motor response specification that involves time-based correction and the attenuation of sensori-motor noise. This notion of prediction and estimation in the presence of temporal delays and sensori-motor noise is the cornerstone of theoretical accounts of tracking and continuous manual control.

The forerunner among theoretical accounts of tracking performance is the optimal control model (OCM, see e.g., Baron & Levinson, 1980; Baron, 1984). Like the *TOTE* model of Miller, Galanter, and Pribram (1960), the optimal control model posits that the human operator acts as a servo-mechanism and responds so as to minimize the difference between displayed current system status and the ideal goal state for the system (cf., Sheridan & Ferrell, 1974). In a dynamic system, observed system status becomes misaligned with the goal state due to disturbances. For, example in guiding an aircraft down a glide path, wind can cause disturbances in the trajectory of motion thereby requiring the pilot to make control adjustments to realign the aircraft's altitude and attitude. In complex systems, such as aircraft, the totality of dynamic operational characteristics can be conveyed in a state space representation of the system. The spatio-temporal characteristics of certain elements of this state space are conveyed to the operator through a dynamic display or through direct perception of the system environment. While most system displays are visual, some may be auditory, and some may be combined audio-visual displays. Similarly, in direct perception of the system environment, control status may be conveyed through visual, auditory, or combined visual and auditory cues regarding spatial layout and spatial dynamics. In any case, the operator perceives these system states through sensory computation in order to develop a common spatio-temporal representation of system dynamics (see Figure 21). These computational sensory processes require a variable amount of time, thus introducing a temporal lag in the operators spatio-temporal representation of the system. Furthermore, sensory processes necessarily introduce noise into the operators spatio-temporal representation. Consequently, time-based correction of the inputs as well as filtering of the sensory noise must be carried out. Moreover, this time-based correction and noise filtering must also compensate for delays and noise that will be subsequently introduced in the course of planning and emitting the responsive motor action. In Figure 21, these functional processes are encapsulated in the time-based corrector and noise filter. In the complete specification of OCM, these processes are presumed to be carried out by a Kalman estimator and predictor whose purpose is to "...generate the best estimate of the current state of the system variables, based on the noisy, delayed perceptual information available" (Baron & Levinson, 1980, p. 91). This estimation process serves to correct for the stochastic variations in spatial and temporal uncertainties thereby producing a predictive representation of the future system state at the instant of control action. Based upon this predictive representation, the motor response program is optimized by a weighting of the control gain parameters in reference to the predicted system state (Baron & Levinson, 1980; Wickens, 1984). That is, response force parameters can be represented in a gain matrix that reflect the estimated deviation between the system state and goal state at the instant the response is made. This gain matrix accounts for the estimated spatio-temporal characteristics of the system as well as the relationships between the system and its displays and controls. Thus, compensation of sensori-motor delays, information processing noise, and system display and control gains is achieved through filtering, time-based correction, and optimal force parameter weighting.

This scheme presented in the optimal control model, when considered discretely, can be equated with motor response planning and programming considered earlier in reference to rapid aimed movements (cf. Wickens, 1984). Indeed a singular response in reference to a system disturbance can be sufficiently described by such action plans and motor programs. However, in a dynamic system where state parameters are in continual flux, a closed-loop sequence of perceptual-motor integration must occur to respond to the inherent instability of the system with respect to its goal state. Consequently, feedback regarding the system state after the response execution must be provided. This feedback is necessary for two reasons. First, responses are parameterized based upon stochastical estimations of noise and delays in processing and therefore exhibit inherent variance that must subsequently be minimized through successive approximations. Second, in the course of information processing and response execution, a highly dynamic system will be susceptible to further sources of disturbance not compensated for in the estimation process. These unanticipated disturbances increase the separation between desired system states and the actual immediate state. Variability in response outputs relative to system status will therefore be exacerbated. This variability can be reduced by successive iterations through the sensori-motor control loop. That is, after an initial ballistic response action, movement comes to be guided through feedback guided current control (cf., Wickens, 1984).

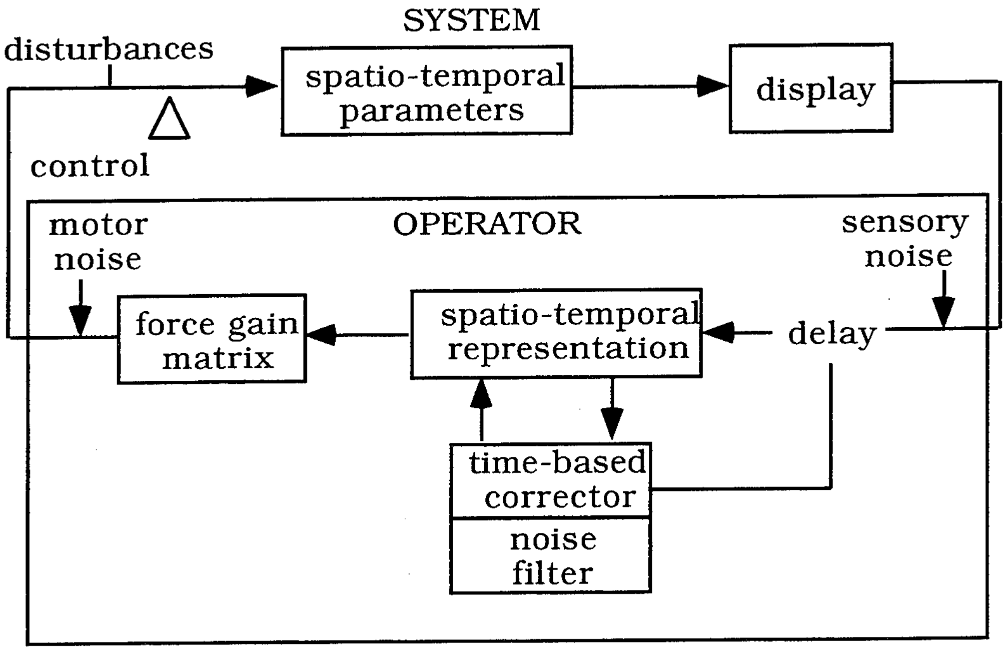

Figure 21. A modified version of the optimal control model demonstrating dynamic spatial and temporal parameterization of control outputs coupled with time based correction of sensori-motor delays (based on Baron & Levinson, 1980; Wickens, 1984).

Feedback guided current control of a system is characterized by the updating of spatio-temporal parameters for responsive motor programming through the processing of information from various spatial modalities such as vision and audition. From this sensory information, predictions must be made to compensate for temporal delays that intervene between the time when information is received and the time when a updated response occurs. Consequently, the response parameter updating that occurs during feedback must include time-based correction. Moreover, it must estimate the variability in the signal and the variability of sensori-motor processing in order to attenuate noise effects in the specification of response parameters. The uncertainty associated with signal transmittal in the course of these feedback loops demonstrates the importance of display quality in conveying information regarding the system state at the time of response execution relative to the state of control effectors. Thus, the accuracy of force parameter specification in continually updated motor plans carried out in tracking tasks is highly dependent upon the quality of spatio-temporal information that can be obtained from system displays.

The quality of information provided by system displays is highly dependent on its ability to convey predictive information regarding the state of the system at the time of the motor response. Prediction of future states can be perceptually conveyed by highlighting higher order spatio-temporal regularities in the fluctuation of system state parameters. For example, a pursuit display, in which the direction and velocity of a spatially represented system parameter is displayed, conveys more information than a compensatory display in which changes in spatially represented system parameters are inherently intertwined with changes in control parameters. That is, in a pursuit display, the spatio-temporal change in system states and the spatio-temporal change produced by the control response are separately conveyed. In this manner, the operator can independently observe the changes in system states caused by external disturbances and those changes caused by control actions. In compensatory displays, on the other hand, only the relative difference between actual and desired states is conveyed, consequently confounding those differences caused by dynamic fluctuations in the system and those caused by motor responses made by the operator. Although more information is conveyed through pursuit displays as compared to compensatory displays, performance differences between these alternative display modes is a function of the operator's ability to perceptually monitor multiple spatial representations independently. Therefore, while tracking with pursuit displays demonstrates superior performance at low to intermediate rates of change in the spatio-temporal characteristics of system displays, at high rates of flux, performance using compensatory displays demonstrates superiority (see Pew, 1974). Thus, information quality is dependent upon the characteristics of the system and the sensory abilities of the perceiver. However, these limitations in the perceptual and cognitive abilities of the operator can be ameliorated through display augmentation.

Poulton (1974) describes four forms of display augmentation: (1) rate augmented displays, (2) quickened displays, (3) predictor displays, and (4) preview displays (cf., Sheridan & Ferrell, 1974; Wickens, 1984). Rate augmentation involves the direct presentation of higher-order rate information and is particularly beneficial in controlling systems whose temporal fluctuations are highly dynamic and whose control device inputs are first-order or higher. An example of a rate augmented display is a pursuit display in which velocity vectors of the system states are conveyed along with instantaneous position information. Quickened displays, on the other hand, provide information regarding predicted future system states in lieu of presentation of information regarding current system states. These displays are distinguished from predictor displays that convey both the present states of the system parameters and the future predicted states of these parameters concurrently (see Sanders & McCormick, 1993). Thus, in both quickened and predictor displays, information is made available that can directly compensate for system lags or delays between the initiation of the response and its impact on the system. Finally, preview displays provide information regarding the future spatial path of a represented system parameter thereby providing information regarding the spatial movement requirements of future controlling actions. Displays of dynamic system parameters can be augmented through various combinations of rate, prediction, and preview information. The degree to which this augmentation improves tracking abilities interacts with the predictability of future system states (cf., Pew, 1974; Folds, Gerth & Engelman, 1987). That is, greater amounts of augmentation are required when the instantaneous predictability of a future system state is low due to the fact that perturbations introduced to the system are highly variable or random in nature. On the other hand, when disturbances to the system are regular and predictable, tracking performance is less dependent upon augmentation in the form of preview or enhanced predictive estimations of future system states. Thus, in highly dynamic and unpredictable environments, continuous control of system state variables is influenced by the quality of information provided to the operator. This quality of information is impacted not only by the degree and kind of display augmentation utilized, but also by the modality through which system state information is presented.

While the majority of tracking displays are presented visually due to the superior nature of visual spatial abilities, tracking of aurally presented system state representations has been demonstrated to be quite accurate albeit not at the level of performance attainable with visual displays (see Poulton, 1974). Moreover, several attempts have been made to incorporate auditory tracking displays into dynamic control scenarios where there are task imposed limitations on conveying information visually or where operator disabilities prohibit the individual from performing the task visually. For example, early work on maneuvering an aircraft by auditory spatial reference conveyed through variations in intensity, frequency, and temporal patterning - the FLYBAR (FLYing By Auditory Reference) project - demonstrated that humans are capable of learning the mappings between spatial deviations conveyed through auditory cues and emit appropriate tracking responses (see Poulton, 1974, Wenzel, 1991 for detailed accounts). Later efforts exploring auditory tracking employed lateralized auditory cues in compensatory displays that signaled the size and direction of spatial error in tracking (see Poulton, 1974). Essentially all past implementations of auditory displays for tracking were compensatory in nature and consequently suffered from the confounding of system disturbances and control inputs as well as the inability to simultaneously convey predictive and current system states. Consequently, application of these auditory compensatory displays has been quite limited. The advent of three-dimensional virtual auditory displays has created great promise for aurally conveying spatial parameters in dynamic systems (see Wenzel, 1991 for an overview). However, due to the novelty of these systems, implementation and empirical observation of tracking performance using this localized auditory spatial information has not yet been a focus of research. Nonetheless, the use of auditory information in the form of frequency, temporal pattern, or intensity changes has been employed as redundant information to visually conveyed spatial parameters. By incorporating both visual and auditory display information, tracking abilities can be augmented through a provision for redundancy gain. Wickens (1984) suggests that in bimodal presentation of system state information, redundant error information presented in the auditory modality aids performance. Therefore, visual pursuit displays coupled with auditory compensatory displays appear to be capable of producing redundancy gain in the transmittal of dynamic spatial information regarding system states and consequently produce superior tracking performance. Therefore, the ability to convey spatio-temporal parameters necessary for defining control requirements in tracking tasks can be enhanced by multimodal presentations of system dynamics. The quality of control responding in relation to these parameters is fundamentally dependent upon the ability of the operator to detect changes and make predictive judgments regarding the properties of the tracking display.

These theoretical accounts and empirical descriptions of continuous manual control can be further extended to consideration of complex tasks involving translation across spatial dimensions over time. For example, continuous manual control is particularly amenable to descriptions of human-machine interactions in transportation systems. The perceptual guidance of motor actions in these scenarios was a particular focus of the ecological approach to perception developed by James Gibson (see Gibson, 1979) that was discussed in detail at the outset of this paper. Thus, in further discussion a reconsideration of this ecological perspective in reference to the motor guidance of locomotor activities will serve to demonstrate the complex coordination between sensation and action and integrate theoretical accounts of sensory information processing and motor response planning.

Locomotion

When an organism traverses across the terrain, either of its own accord or with the aid of a transportation vehicle, global changes in the optical array occur. This activity exemplifies the nature of sensori-motor coordination as emphasized by Dewey's (1896, in Watson, 1979, p. 239) comment that "the so-called response is not merely to the stimulus, it is *into* it". That is, in the course of locomotion, relationships between spatial features are highly dynamic and therefore can only be recovered through consideration of higher order spatio-temporal invariants. Earlier, it was demonstrated that the critical invariants for specifying parameters for locomotor guidance are the point of optical expansion (POE) and the relative rate of optical expansion (tau - τ). Therefore, these invariants are the critical parameters for specifying motor guidance plans for successful navigation through a cluttered terrain or along a specified path. For example, in specifying a dynamic action plan for operating a motor vehicle, the relationship between optical flow lines emanating from the point of expansion and the course of the road ahead specifies the instantaneous difference between the actual and desired direction of travel (Lee, 1980; Bruce & Green, 1985; Mestre, 1992). Consequently, the parameterization of a steering response plan is dependent upon specification of the point of optical expansion relative to features of the terrain. The instantaneous time-to-contact (τ), on the other hand, specifies the conditions for deceleration and braking in response to stationary and moving obstacles on the road ahead such as other vehicles. In fact, Lee (1980, cf., Bruce & Green, 1985) demonstrate that a driver's current rate of deceleration (D) is adequate, if and only if, the distance required to stop is less than the distance of the obstacle ahead, which is true when:

$$\frac{V(t)^2}{2D} \le Z(t)$$

or when (13)

$$\frac{Z(t)D}{V(t)^2} \ge \tfrac{1}{2}$$

where $V(t) =$ relative velocity at time t and

$Z(t) =$ instantaneous distance at time t

since tau $(\tau) = \dfrac{Z(t)}{V(t)}$

(viz. Equation 2)

deceleration is adequate iff (14)

$$\frac{d\tau}{dt} \le -\tfrac{1}{2}$$

The specification of such parameters for motor action demonstrate that the time-sampled rate of dilation of features in the optical array serves as a sufficient visual parameter for the initiation of locomotor activities in driving tasks such braking and timing turns at intersections in the face of oncoming traffic (see Lee, 1980; Caird & Hancock, 1992).

The continual updating of guidance parameters specified by these higher order invariants can be accounted for through the optimal control model of dynamic response planning and execution. That is, the execution of steering and deceleration responses must be made continually within a feedback loop that includes time-based correction and attenuation of sensori-motor noise introduced in the processing of these invariant spatio-temporal parameters. Thus, locomotion can be seen as a complex interaction of tracking along two dimensions specified by the fronto-parallel and depth planes. Incongruities between actual and desired states along these dimensions are specified through spatio-temporal invariants, namely the point of optical expansion (POE) and the relative rate of optical expansion (τ). Action plans for motor guidance during locomotion consequently can be sufficiently specified through these parameters and continually updated to minimize discrepancies between the actual and desired position and course of travel relative to objects in the organism's path. Earlier, it was demonstrated that these higher order spatio-temporal invariants are capable of being conveyed through both the auditory and visual modalities. These multimodally determined spatio-temporal invariants can consequently serve as parameters for motor guidance plans that the organism initiates and continually updates in order to successfully traverse the terrain of its surrounding environment and emit motor acts critical for its survival. This coordination of multimodal sensory inputs and control actions in the presence of continual flux in the optical structure involves a complex information-processing system that is specifically attuned to perceiving these higher order spatio-temporal parameters and implementing them in responsive motor guidance plans. In the following chapter, a comprehensive model will be developed to account for the processes through which information regarding the dynamic spatial layout can be incrementally accrued from the auditory and visual modalities and subsequently serve to specify the spatio-temporal parameters necessary for coordinating motor actions.

AN INTEGRATED THEORY OF SPATIO-TEMPORAL DYNAMICS

Throughout this report, the integration of ecological and computational paradigms for describing perceptual-motor coordination has been emphasized. Through the ecological approach, the spatio-temporal properties or invariants in the environment critical for perception have been delineated. Computational accounts have suggested how the specific structuring and coding of information in sensory modalities can serve to analyze and synthesize a spatio-temporal representation of the environment. It has further been demonstrated that, despite the differential characteristics of information analyzed by the visual and auditory modalities, they converge upon a unified, common representation of the dynamic spatial layout. Based upon this common representation of spatio-temporal information, goal directed motor actions can be specified and emitted by a dynamic perceiving organism. These motor actions inherently change the structure of the dynamic spatial layout, and this change is subsequently conveyed through an updated spatio-temporal representation (see Figure 22).

The nature of this coordinative relationship between the environment and the dynamic perceiver is reflected in the theoretical issues and empirical findings regarding visual and auditory spatial perception and responsive motor action detailed above. Several general principles can be extracted from this consideration and integrated into a comprehensive model of perceptual-motor coordination. Foremost in the construction of this model of perceptual motor coordination is the fact that the critical features of the environment for perception and action are objects and events that are informationally defined in terms of their spatio-temporal properties. Thus, the role of perceptual systems is to computationally synthesize these spatio-temporal properties into a coherent representation which can subsequently serve to parameterize responsive motor plans and actions. However, this synthesis is necessarily preceded by an analysis of the light and sound impinging upon the visual and auditory sense organs. The analysis of these transduced light and sound signals by the visual and auditory system allows for the incremental accumulation of information regarding spatial dynamics and spatial layout based upon the functional requirements of the organism.

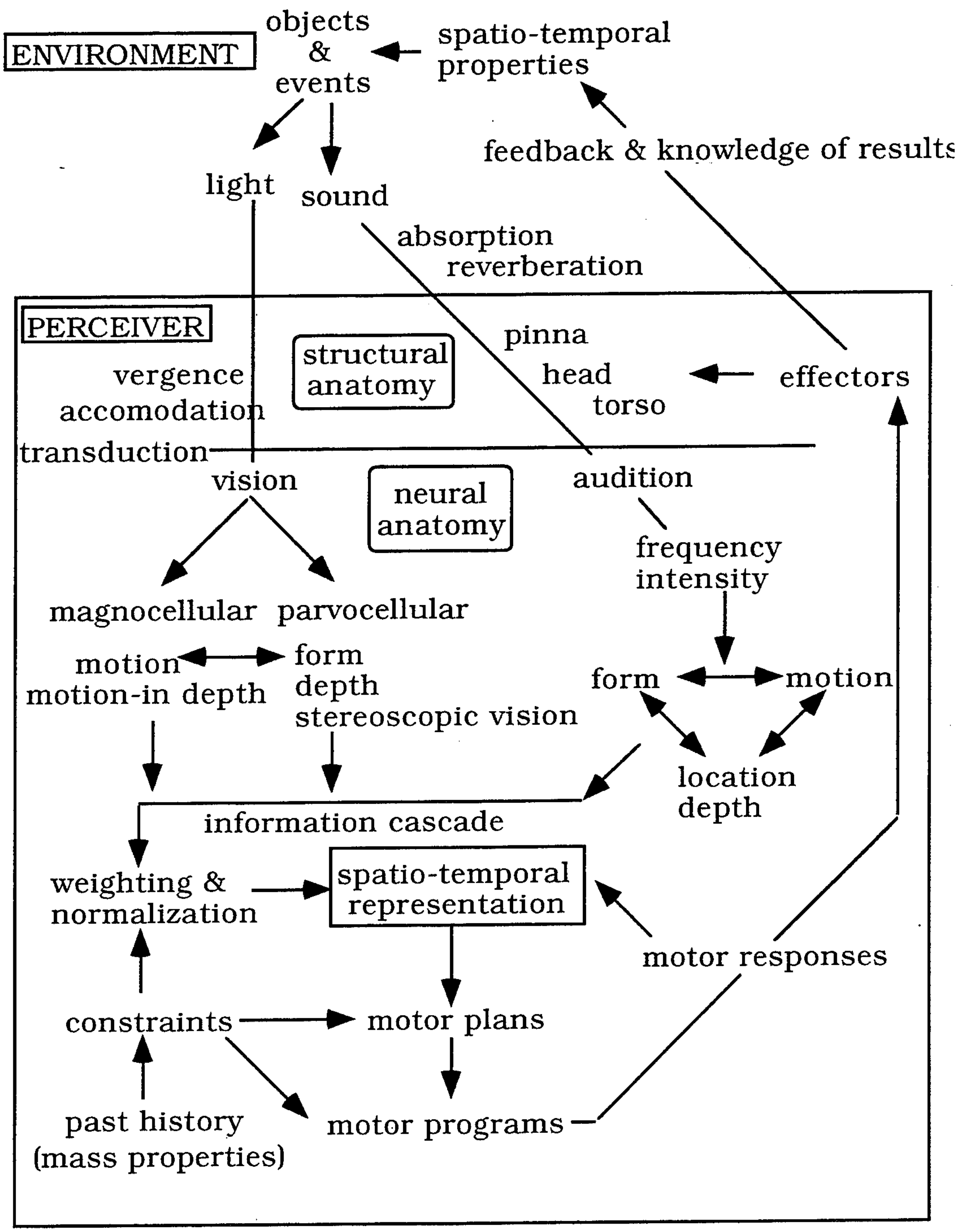

Figure 22. A schematic description of the coordination between dynamic visual and auditory spatial percepts and motor actions.

This incremental accrual of dynamic spatial information is particularly evident in the visual system where transient motion characteristics are built up rapidly through "short-range" processes conveying general motion in the fronto-parallel plane and in depth largely independent of extensive form processing. The subsequent interaction of this rapid "short-range" motion analysis carried out predominantly by the magnocellular system with fine detailed analysis of the spatial layout carried out by the parvocellular system allows for a "long-range" specification of the dynamic spatial layout that accrues over time. Therefore, the spatio-temporal representation of the environment is the product of a cascading flow of information from rapid motion detecting mechanisms and slower form detecting mechanisms. The inherent interaction of form and motion information is subject to the imposition of constraints and the weighting of information parameters that reflects the prior history of the perceiver. For example, prior experience with grasping distant objects can lead to the weighting of monocular and binocular motion and depth cues used to construct a spatio-temporal representation of the three-dimensional scene. Thus, in the visual system, the early, separate analysis of motion and form information is incrementally synthesized into a unique spatio-temporal representation of the ambient optic array.

In the auditory system, such a distinctive separation between motion and form information is not possible because acoustical analysis of form and spatial layout is highly interdependent. Nonetheless, the auditory modality can provide important spatial cues not perceptible through the visual modality due to the restricted field of view. These auditory spatial cues, like their visual counterparts, are important for parameterizing motor plans and guiding motor actions. Auditory spatial cues are particularly critical for specifying plans for orientation responses and the regulation of head and eye movements to acquire spatial targets at high degrees of eccentricity. In this manner, the auditory system can provide rapid rudimentary spatio-temporal cues that can direct viewing and subsequently allow for a more extensive build-up of information regarding the dynamic spatial layout. In this manner, auditory and visual cues can lead to the incremental accrual of spatio-temporal properties regarding objects in the environment.

Thus, the current model emphasizes the common representation of spatio-temporal parameters derived from the auditory and visual modality. Moreover, this model stresses that this dynamic spatial information is incrementally accrued in a cascading fashion. Furthermore, it is posited that this common spatio-temporal information array contains information about object position and object dynamics as well as information regarding the dynamics of the perceiver in relation to these objects. Therefore, this representational information array contains higher order spatio-temporal invariants in addition to information conveying static spatial layout. This representation operates to convey these higher order spatio-temporal parameters to motor effectors in the form of plans for action. The need for rapid and precise parameterization for these motor plans is realized through this incremental cascading scheme for the acquisition of multimodal spatio-temporal information. That is, low level motion and gross form information allows for rapid initiation of motor responses, whereas higher level information regarding detailed form and location can subsequently be accrued and relayed to ongoing motor actions thereby allowing for precise outcomes. In this manner, a complex interaction between the multiple modalities, pathways, and computational processes of perception can converge upon a unified array of spatio-temporal attributes in an incremental, cascading fashion. Coordination between these inputs and responsive motor actions consequently can be achieved in a parsimonious and economical fashion. Motor plans can be initiated with speed and accuracy inside this coordination and can afford the organism the opportunity for interaction with a highly dynamic environment.

The particularly striking feature of this model is its description of the build-up of spatio-temporal information across modalities. Based upon this multimodal information accrual process, it is predicted that spatio-temporal information accrued in one modality can aid the planning of motor responses even when these actions must ultimately be executed in relation to objects perceived by another modality. This is particularly true of auditory spatio-temporal cues that can be perceived beyond the field of view thereby providing preview for visually guided responses. Ongoing research that I began in 1993 is focusing on this common spatio-temporal representation of auditory and visual percepts and the potential benefits and limitations of augmenting visual displays with dynamic auditory preview information. Preliminary findings from this work (see Elias, 1994) indicate great potential for the future application of spatial auditory displays for augmenting visual cues in complex task domains such as aviation. These findings suggest that dynamic spatial cues conveyed aurally and visually can be commonly represented in formulating action plans for motor responses. Consequently, dynamic spatial auditory cues can be utilized as preview for subsequent visually guided responses.

This preliminary research demonstrated that both velocity and trajectory information can be commonly conveyed across modalities thereby enabling dynamic auditory cues to serve as preview for estimating the position and speed of impending visual targets (see Elias, 1994).

However, to date, no research has been done on the effects of dynamic auditory preview cues conveying target velocity and trajectory information on visual target identification. Such research is needed to assess the functional utility of three-dimensional auditory displays in highly dynamic tasks environments where visual task demands and restrictions in the field of view limit visual monitoring capabilities. Current research being conducted under the Armstrong Laboratory In-house Laboratory Independent Research (ILIR) program is addressing this very issue. The results of this research will serve to answer important questions regarding the degree and manner in which dynamic spatial auditory cues can be utilized for enhancing visual displays. Application domains such as tactical displays, airborne situational displays, and air traffic control displays are highly dynamic in nature and must convey critical information regarding object speed and trajectory. Findings from this research program will serve to identify task dependent criteria for assessing the utility and potential limitations of dynamic auditory preview displays for visual target identification in these domains of application.

REFERENCES

1. Abrams, R. A., Coordination of eye and hand for aimed limb movements. In (L. Procteau & D. Elliott, Eds.) *Vision and Motor Control* (pp. 129-152). New York, NY: Elsevier, 1992.

2. Abrams, R. A., Meyer, D. E., & Kornblum, S., Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the ocular system. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 529-543, 1989.

3. Adams, J. A., Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin, 101*(1), 41-74, 1987.

4. Adelson, E. H. & Bergen, J. R., Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America, 2* (2), 284-299, 1985.

5. Anstis, S., Motion perception in the frontal plane. In (K. R. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Human Perception and Performance.* New York, NY: John Wiley and Sons, 1986.

6. Arditi, A., Binocular vision. In (K. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Perception and Human Performance: Volume I, Sensory Processes and Perception* (Ch. 23). New York, NY: Wiley, 1986.

7. Auerbach, C. & Sperling, P., A common auditory-visual space: Evidence for its reality. *Perception and Psychophysics, 16*(1), 129-135, 1974.

8. Baron, S., A control theoretic approach to modelling human supervisory control of dynamic systems. In (W. B. Rouse, Ed.) *Advances in Man-Machine Systems Research, Vol. 1* (pp. 1-47). JAI Press, 1984.

9. Baron, S. & Levinson, W. H., The optimal control model: Status and future directions. *IEEE Proceedings,* 1980.

10. Batteau, D. W., The role of the pinna in human localization. *Proceedings of the Royal Society of London, Series B, 168,* 158-180, 1967.

11. Begault, D. R., Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation. *Human Factors, 35*(4), 707-717, 1993.

12. Begault, D. R. & Wenzel, E. M., Headphone localization of speech stimuli. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 82-85). Santa Monica, CA: Human Factors Society, 1991.

13. Biederman, I., Recognition by components: A theory of human image understanding. *Psychological Review, 94*(2), 115-14, 1987.

14. Bootsma, R. J. & Peper, C. E., Predictive visual information sources for the regulation of action with special emphasis on catching and hitting. In (L. Procteau & D. Elliott, Eds.) *Vision and Motor Control* (pp. 285-314). New York, NY: Elsevier, 1992.

15. Boulton, J. C. & Hess, R. F., Luminance contrast and motion detection. *Vision Research, 30*(1), 175-179, 1990a.

16. Boulton, J. C. & Hess, R. F., The optimal displacement for the detection of motion. *Vision Research, 30*(7), 1101-1106, 1990b.

17. Braddick, O. J., A short range process in apparent motion. *Vision Research, 14*, 519-527, 1974.

18. Braddick, O. J., Low-level and high-level processes in apparent motion. *Philosophical Transactions of the Royal Society of London, Series B, 290*, 137-151, 1980.

19. Braunstein, M. L., Andersen, G. J., Rouse, M. W., & Tittle, J. S., Recovering viewer-centered depth from disparity, occlusion, and velocity gradients. *Perception and Psychophysics, 40*(4), 216-224, 1986.

20. Brown, A. E. & Hopkins, H. K., Interaction of the auditory and visual sensory modalities. *Journal of the Acoustical Society of America, 41*(1), 1-6, 1967.

21. Bruce, V. & Green, P., *Visual Perception: Physiology, Psychology and Ecology.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1985.

22. Butler, R. A., An analysis of the monaural displacement of sound in space. *Perception and Psychophysics, 41*(1), 1-7, 1987.

23. Caird, J. K. & Hancock, P. A., Perception of oncoming vehicle time-to-arrival. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1378-1382). Santa Monica, CA: Human Factors Society, 1992.

24. Cavanagh, P., Reconstructing the third dimension: Interactions between color, texture, motion, binocular disparity, and shape. *Computer Vision, Graphics and Image Processing, 37*, 171-195, 1987.

25. Coleman, P. D., Failure to localize the source distance of an unfamiliar sound. *Journal of the Acoustical Society of America, 34*(3), 345-346, 1962.

26. Coleman, P. D., An analysis of cues to auditory depth perception in free space. *Psychological Bulletin, 60*(3), 302-315, 1963.

27. DeLucia, P. R., Pictorial and motion-based information for depth perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*(3), 738-748, 1991.

28. Deutsch, D., Musical illusions. *Scientific American, 233*(4), 92-104, 1975.

29. Deutsch, D., Auditory pattern recognition. In (K. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Perception and Human Performance: Volume II, Cognitive Processes and Performance* (Ch. 32). New York, NY: Wiley, 1986.

30. DeValois, R. L. & DeValois, K. K., Spatial vision. *Annual Review of Psychology, 31,* 309-341, 1980.

31. DeValois, R. L. & DeValois, K. K., *Spatial Vision.* New York, NY: Oxford University Press, 1988.

32. Dewey, J., The reflex arc concept in psychology. *Psychological Review, 3,* 357-370. Reprinted in R. I. Watson (1979), *Basic Writings in the History of Psychology.* New York, NY: Oxford University Press, 1896.

33. Doll, T. J. & Hanna, T. E., Enhanced detection with bimodal sonar displays. *Human Factors, 31*(5), 539-550, 1989.

34. Doll, T. J., Hanna, T. E., & Russotti, J. S., Masking in three-dimensional auditory displays. *Human Factors, 34*(3), 255-26, 1992.

35. Elias, B., *Aural preview for visually guided target acquisition and aiming.* Doctoral Thesis: Georgia Institute of Technology, Atlanta, GA, 1994.

36. Ellis, S. R., Tyler, M., Kim, W. S., & Stark, L., Three dimensional tracking between display and control axes. *Proceedings of the SAE International Conference on Environmental Systems*, 1991.

37. Fidell, S., Sensory function in multimodal signal detection. *Journal of the Acoustical Society of America, 47*(4), 1009-1015, 1970.

38. Folds, D. J., Advanced audio displays in aerospace systems: Technology requirements and expected benefits. *NAECON '90*, 739-743, 1990.

39. Folds, D. J., Gerth, J. M., & Engelman, W. R., Enhancement of human performance in manual target acquisition and tracking. *USAFSAM-TR-86-18.* Brooks AFB, TX: USAF School of Aerospace Medicine, ,1987.

40. Gaver, W. W., Smith, R. B., & O'Shea, T., Effective sounds in complex systems: The Arkola simulation. *Proceedings of the Conference on Computer Human Interaction (CHI) '91* (pp. 85-90). New York, NY: Association for Computing Machinery, Inc, 1991.

41. Gibson, J. J., *The Ecological Approach to Visual Perception.* Boston, MA: Houghton Mifflin, 1979.

42. Gogel, W. C., Relative motion and the adjacency principle. *Quarterly Journal of Experimental Psychology, 26,* 425-437, 1974.

43. Gogel, W. C., The adjacency principle in visual perception. *Scientific American, 238*(5), 126-139, 1978.

44. Goldstein, E. B., *Sensation and Perception (2nd. Ed.).* Belmont, CA: Wadsworth, 1984.

45. Good, M. D. & Gilkey, R. H., Masking between spatially separated sounds. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 253-257). Santa Monica, CA: Human Factors Society, 1992.

46. Handel, S., *Hearing: An Introduction to the Perception of Auditory Events.* Cambridge, MA: MIT Press, 1989.

47. Hess, R. F. & Snowden, R. J., Temporal properties of human visual filters: number, shapes and spatial covariation. *Vision Research, 32*(1), 47-59, 1992.

48. Hildreth, E. C., Grzywacz, N. M., Adelson, E. H., & Inada, V. K., The perceptual buildup of three-dimensional structure from motion. *Perception and Psychophysics, 48*(1), 19-36, 1990.

49. Hochberg, J., Representation of motion and space in video and cinematic displays. In (K. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Perception and Human Performance: Volume I, Sensory Processes and Perception* (Ch. 22). New York, NY: Wiley, 1986.

50. Jagacinski, R. J., & Monk, D. L., Fitts' Law in two dimensions with hand and head movements. *Journal of Motor Behavior, 17,* 77-95, 1985.

51. Jagacinski, R. J., Johnson, W. W., & Miller, R. A., Quantifying the cognitive trajectories of extrapolated movements. *Journal of Human Psychology: Human Perception and Performance, 9*(1), 43-57, 1983.

52. Jagacinski, R. J., Repperger, D. W. , Moran, M. S., & Ward, S. L., The microstructure of rapid discrete movements. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 309-320, 1980.

53. Johansson, G., von Hofsten, C., Jansson, G., Event perception. *Annual Review of Psychological, 31*, 27-63, 1980.

54. Julesz, B., Textons, the elements of texture perception and their interactions. *Nature, 290,* 91-97, 1981.

55. Kim, W. S., Ellis, S. R., Tyler, M. E., Hannaford, B., & Stark, L. W., Quantitative evaluation of perspective and stereoscopic displays in three-axis manual tracking tasks. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-17*(1), 61-72, 1987.

56. Kinsler, L. E., Frey, A. R., Coppens, A. B., & Sanders, J. V., *Fundamentals of Acoustics (3rd Ed.).* New York, NY: John Wiley & Sons, 1982.

57. Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M., Measurement and modeling of depth cue combination: In defense of weak fusion. In *Mathematical Studies in Perception and Cognition 91-3*, 1991.

58. Larish, J. F. & Flach, J. M., Sources of optical information useful for perception of speed of rectilinear self-motion. *Journal of Experimental Psychology: Human Perception and Performance, 16*(2), 295-302, 1990.

59. Lee, D. N., The optic flow field: The foundation of vision. *Philosophical Transactions of the Royal Society of London, Series B, 290,* 169-179, 1980.

60. Lee, D. N., Young, D. S., Reddish, P. E., Lough, S., & Clayton, T. M. H., Visual timing in hitting an accelerating ball. *Quarterly Journal of Experimental Psychology, 35A,* 333-346, 1983.

61. Lennie, P., Trevarthen, C., Van Essen, D., & Wassle, H., Parallel processing of visual information. In (L. Spillman & J. S. Werner, Eds.) *Visual Perception: The Neurophysiological Foundations* (pp. 103-128). San Diego, CA: Academic Press, 1990.

62. Livingstone, M. & Hubel, D., Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science, 240,* 740-749, 1988.

63. Loomis, J. M., Herbert, C., & Cicinelli, J. G., Active localization of virtual sounds. *Journal of the Acoustical Society of America, 88*(4), 1757-1764, 1990.

64. Loveless, N. E., Brebner, J., & Hamilton, P., Bisensory presentation of information. *Psychological Bulletin, 73*(3), 161-199, 1970.

65. Makous, J. C., & Middlebrooks, J. C., Two-dimensional sound localization by human listeners. *Journal of the Acoustical Society of America, 87*(5), 2188-2200, 1990.

66. Mack, A., Perceptual aspects of motion in the frontal plane. In (K. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Perception and Human Performance: Volume I, Sensory Processes and Perception* (Ch. 17). New York, NY: Wiley, 1986.

67. Marr, D. & Hildreth, E., Theory of edge detection. *Proceedings of the Royal Society of London, Series B, 207,* 187-217, 1980.

68. Marr, D. & Nishihara, H. K., Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London, Series B, 200,* 269-294, 1978.

69. Marr, D. & Poggio, T., Cooperative computation of stereo disparity. *Science, 194,* 283-287, 1976.

70. Marr, D. & Poggio, T., A computational theory of human stereo vision. *Proceedings of the Royal Society of London, Series B, 204,* 301-328, 1979.

71. Marr, D., *Vision: A Computational Investigation into Human Representation and Processing of Visual Information.* New York, NY: W. H. Freeman, 1982.

72. Mayhew, J. E. W. & Frisby, J. P., Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence, 17,* 349-385, 1981.

73. Mazur, J. E., *Learning and Behavior.* Englewood Cliffs, NJ: Prentice Hall, 1990.

74. McKinley, R. L., Erickson, M. A., & D'Angelo, W. R., 3-dimensional auditory displays: Development, applications, and performance. *Aviation, Space, and Environmental Medicine, May 1994,* A31-A38, 1994.

75. Mershon, D. H. & Hutson, W. E., Toward the indirect measurement of perceived auditory distance. *Bulletin of the Psychonomic Society, 29*(2), 109-112, 1991.

76. Mershon, D. H. & King, E., Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception and Psychophysics, 18*(6), 409-415, 1975.

77. Mershon, D. H., Ballenger, W. L., Little, A. D., McMurtry, P. L., & Buchanan, J. L., Effects of room reflectance and background noise on perceived auditory distance. *Perception, 18,* 403-416, 1989.

78. Mestre, D. R., Visual perception of self-motion. In (L. Procteau & D. Elliott, Eds.) *Vision and Motor Control* (pp. 421-438). New York, NY: Elsevier, 1992.

79. Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E., & Smith, J. E. K., Optimality in human motor performance: Ideal control of rapid aimed movements. *Psychological Review, 95,* 340-370, 1988.

80. Meyer, D. E., Smith, J. E. K., Abrams, R. A., Kornblum, S., & Wright, C. E., Speed-accuracy tradeoffs in rapid aimed movements: Toward a theory of rapid voluntary action. In M. Jeannerod (Ed.), *Attention and Performance XIV.* (pp. 173-226). Hillsdale, NJ: Erlbaum, 1990.

81. Middlebrooks, J. C. & Green, D. M., Sound localization by human listeners. *Annual Review of Psychology, 42,* 135-159, 1991.

82. Miller, G. A., Galanter, E., & Pribram, K. H., *Plans and the Structure of Behavior.* New York, NY: Holt-Dryden, 1960.

83. Moore, B. C. J., *An Introduction to the Psychology of Hearing (2nd Ed.).* San Diego, CA: Academic Press, 1988.

84. O'Leary, A. & Rhodes, G., Cross-modal effects on visual and auditory object perception. *Perception & Psychophysics, 35*(6), 565-569, 1984.

85. Ohzawa, I, DeAngelis, G. C., & Freeman, R. D., Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science, 249,* 1037-1041, 1990.

86. Osborn, W. C., Sheldon, R. W., & Baker, R. A., Vigilance performance under conditions of redundant and nonredundant signal presentation. *Journal of Applied Psychology, 47*(2), 130-134, ,1963.

87. Patterson, R. & Martin, W. L., Human stereopsis. *Human Factors, 34*(6), 669-672, 1992.

88. Perrott, D. R., Buck, V., Waugh, W., & Strybel, T. Z., Dynamic auditory localization: Systematic replication of the auditory velocity function. *The Journal of Auditory Research, 19,* 277-285, 1979.

89. Perrott, D. R. & Marlborough, K., Minimum audible movement angle: Marking the end points of the path traveled by a moving sound source. *Journal of the Acoustical Society of America, 85*(4), 1773-1775, 1989.

90. Perrott, D. R., Saberi, K., Brown, K., & Strybel, T. Z., Auditory psychomotor coordination and visual search performance. *Perception and Psychophysics, 48*(3), 214-226, 1990.

91. Perrott, D. R., Sadralodabai, T., Saberi, K., & Strybel, T. Z., Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. *Human Factors, 33*(4), 389-400, 1991.

92. Petersik, J. T., The two-process distinction in apparent motion. *Psychological Bulletin, 106*(1), 107-127, 1989.

93. Pew, R. W., Human perceptual-motor performance. In (B. Kantowitz, Ed.) *Human Information Processing: Tutorials in Performance and Cognition.* New York: Wiley, 1974.

94. Pew, R. W., & Rosenbaum, D. A., Human movement control: Computation, representation, and implementation. In R. C. Atkinson, R. J. Hernstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' Handbook of Experimental Psychology,* 2nd Edition (pp. 473-509). New York: Wiley, 1988.

95. Platt, B. B. & Warren, D. H., Auditory localization: The importance of eye movements and a textured environment. *Perception and Psychophysics, 12*(2B), 245-248, 1972.

96. Poggio, G. F. & Poggio, T., The analysis of stereopsis. *Annual Review of Neuroscience, 7,* 379-412, 1984.

97. Posner, M. I., *Chronometric Explorations of Mind.* New York, NY: Oxford University Press,. 1986.

98. Poulton, E. C., *Tracking Skill and Manual Control.* New York, NY: Academic Press, 1974.

99. Rayleigh, Lord, On our perception of sound direction. *The Philosophical Magazine, 13,* 214-232, 1907.

100. Regan, D. & Beverley, K. I., Visually guided locomotion: Psychophysical evidence for a neural mechanism sensitive to flow patterns. *Science, 205*, 311-313, 1979.

101. Regan, D. & Beverley, K. I., Visual responses to changing size and to sideways motion for different directions of motion in depth: Linearization of visual responses. *Journal of the Optical Society of America, 70*(11), 1289-1296, 1980.

102. Regan, D. Beverley, K., & Cynader, M., The visual perception of motion in depth. *Scientific American, 241,* 136-151, 1979.

103. Regan, D. M., Kaufman, L, & Lincoln, J., Motion in depth and visual acceleration. In (K. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Perception and Human Performance: Volume I, Sensory Processes and Perception* (Ch. 19). New York, NY: Wiley, 1986.

104. Regan, D., Frisby, J. P., Poggio, G. F., Schor, C. M., & Tyler, C. W., The perception of stereodepth and stereomotion: Cortical mechanisms. In (L. Spillman & J. S. Werner, Eds.) *Visual Perception: The Neurophysiological Foundations* (pp. 317-347) San Diego, CA: Academic Press, 1990.

105. Regan, D., Hamstra, S. & Kaushal, S., Visual factors in the avoidance of front-to-rear-end highway collisions. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1006-1010). Santa Monica, CA: Human Factors Society, 1992.

106. Rosenbaum, D. A., Perception and extrapolation of velocity and acceleration. *Journal of Experimental Psychology: Human Perception and Performance, 1*(4), 395-403, 1975.

107. Salmoni, A. W., Schmidt, R. A., & Walter, C. B., Knowledge of results and motor learning: A review and critical appraisal. *Psychological Bulletin, 95,* 355-386, 1984.

108. Sanders, M. S. & McCormick, E. J., *Human Factors in Engineering and Design (7th Ed).* New York, NY: McGraw-Hill, 1993.

109. Scharf, B. & Houtsma, A. J. M., Audition II: Loudness, pitch, localization, aural distortion, pathology. In (K. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Perception and Human Performance: Volume I, Sensory Processes and Perception* (Ch. 15). New York, NY: Wiley, 1986.

110. Schiff, W. & Oldak, R., Accuracy of judging time to arrival: Effects of modality, trajectory, and gender. *Journal of Experimental Psychology, Human Perception and Performance, 16*(2), 303-316, 1990.

111. Schmidt, R. A., A schema theory of discrete motor skill learning. *Psychological Review, 82,* 225-260, 1975.

112. Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C., Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15,* 352-359, 1989.

113. Schmidt, R. A., Zelasnik, H., Hawkins, B., Frank, J. S., & Quinn, J. T., Motor output variability: A theory for the accuracy of rapid motor acts. *Psychological Review, 86,* 415-451, 1979.

114. Sedgwick, H. A., Space perception. In (K. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Perception and Human Performance: Volume I, Sensory Processes and Perception* (Ch. 21). New York, NY: Wiley, 1986.

115. Sekuler, R., Anstis, S., Braddick, O. J., Brandt, T., Movshon, J. A., & Orban, G., The perception of motion. In (L. Spillman & J. S. Werner, Eds.) *Visual Perception: The Neurophysiological Foundations* (pp. 205-230). San Diego, CA: Academic Press, 1990.

116. Shaw, B. K., McGowan, R. S., & Turvey, M. T., An acoustic variable specifying time-to-contact. *Ecological Psychology, 3*(3), 253-261, 1991.

117. Sheridan, T. B. & Ferrell, W. R., *Man-Machine Systems: Information, Control, and Decision Models of Human Performance.* Cambridge, MA: MIT Press, 1974.

118. Silverman, M. S., Grosof, D. H., DeValois, R. L., & Elfar, S. D., Spatial frequency organization in primate striate cortex. *Proceedings of the National Academy of Sciences (USA), 86*, 711-715, 1989.

119. Sorkin, R. D., Wightman, F. L., Kistler, D. S., & Elvers, G. C., An exploratory study of the use of movement-correlated cues in an auditory head-up display. *Human Factors, 31*(2), 161-166, 1989.

120. Strybel, T. Z. & Perrott, D. R., Discrimination of relative distances in the auditory modality: The success and failure of the loudness discrimination hypothesis. *Journal of the Acoustical Society of America, 76*(1), 318-320, 1984.

121. Strybel, T. Z., Manligas, C. L., & Perrott, D. R., Minimum audible movement angle as a function of the azimuth and elevation of the source. *Human Factors, 34*(3), 267-275, 1992.

122. Todd, J. T., Visual information about moving objects. *Journal of Experimental Psychology: Human Perception and Performance, 7*(4), 795-810, 1981.

123. Tootell, R. B. H., Silverman, M. S., & DeValois, R. L., Spatial frequency columns in primary visual cortex. *Science, 214,* 813-815, 1981.

124. Ullman, S., The interpretation of structure from motion. *Proceedings of the Royal Society of London, Series B, 203,* 405-426, 1979.

125. Walker, N., Meyer, D. E., & Smelcer, J. B., Spatial and temporal characteristics of rapid cursor-positioning movements with electromechanical mice in human-computer interaction. *Human Factor, 35*(3), 431-458, 1993.

126. Warren, D. H., Intermodality interactions in spatial localization. *Cognitive Psychology, 1,* 114-133, 1970.

127. Warren, W. H., Action modes and laws of control for the visual guidance of action. In (O. G. Meijer & K. Roth, Eds.) *Complex Motor Behaviour: 'The' motor-action controversy* (pp. 339-380). New York, NY: Elsevier (North-Holland), 1988.

128. Watson, A. B., Optimal displacement in apparent motion and quadrature models of motion sensing. *Vision Research, 30*(9), 1389-1393, 1990.

129. Watson, A. B., Ahumada, A. J., & Farrell, J .E., Window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays. *Journal of the Optical Society of America, 3* (3), 300-307, 1986.

130. Waugh, W., Strybel, T. Z., & Perrott, D. R., Perception of moving sounds: Velocity discrimination. *The Journal of Auditory Research, 19,* 103-110, 1979.

131. Welsh, R. B. & Warren, D. H., Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*(3), 638-667, 1980.

132. Welsh, R. B. & Warren, D. H., Intersensory interactions. In (K. Boff, L. Kaufman, & J. P. Thomas, Eds.) *Handbook of Perception and Human Performance: Volume I, Sensory Processes and Perception* (Ch. 25). New York, NY: Wiley, 1986.

133. Wenzel, E. M., Three-dimensional virtual acoustic displays. *NASA Technical Memorandum 103835*. Moffett Field, CA: NASA Ames Research Center, 1991.

134. Wenzel, E. M., Wightman, F. L., & Kistler, D. J., Localization with non-individualized virtual acoustic display cues. *Proceedings of the Conference on Computer Human Interaction (CHI) '91* (pp. 351-359). New York, NY: Association for Computing Machinery, Inc, 1991.

135. Wickens, C. D., *Engineering Psychology and Human Performance*. Columbus, OH: Merrill, 1984.

136. Wickens, C. D., *Engineering Psychology and Human Performance. (2nd. Ed.)* . New York, NY: Harper Collins, 1992.

137. Wicker, F. W., Mapping the intersensory regions of perceptual space. *American Journal of Psychology, 81,* 178-188, 1968.

138. Wightman, F. L. & Kistler, D. J., Headphone simulation of free-field listening. I: Stimulus synthesis. *Journal of the Acoustical Society of America, 85*(2), 858-867, 1989a.

139. Wightman, F. L. & Kistler, D. J., Headphone simulation of free-field listening. II: Psychophysical validation. *Journal of the Acoustical Society of America, 85*(2), 868-878, 1989b.

140. Wilson, H. R., Responses of spatial mechanisms explain hyperacuity. *Vision Research, 26*(3), 453-469, 1986.

141. Wilson, H. R., Levi, D., Maffei, L., Rovamo, J. & DeValois, R., The perception of form: Retina to striate cortex. In (L. Spillman & J. S. Werner, Eds.) *Visual Perception: The Neurophysiological Foundations* (pp. 231-272) San Diego, CA: Academic Press, 1990.